Spending symmetry

Terence Tao

Department of Mathematics, UCLA, Los Angeles, CA 90095
E-mail address: tao@math.ucla.edu

In memory of Garth Gaudry, who set me on the road

Contents

Preface

In February of 2007, I converted my "What's new" web page of research updates into a blog at `terrytao.wordpress.com`. This blog has since grown and evolved to cover a wide variety of mathematical topics, ranging from my own research updates, to lectures and guest posts by other mathematicians, to open problems, to class lecture notes, to expository articles at both basic and advanced levels. In 2010, I also started writing shorter mathematical articles, first on a (now defunct) Google Buzz feed, and now at the Google+ feed

`plus.google.com/114134834346472219368/posts` .

This book collects some selected articles from both my blog and my Buzz and Google+ feeds from 2011, continuing a series of previous books [**Ta2008**], [**Ta2009**], [**Ta2009b**], [**Ta2010**], [**Ta2010b**], [**Ta2011**], [**Ta2011b**], [**Ta2011c**], [**Ta2011d**], [**Ta2012**] based on the blog and Buzz.

The articles here are only loosely connected to each other, although many of them share common themes (such as the titular use of *compactness and contradiction* to connect finitary and infinitary mathematics to each other). I have grouped them loosely by the general area of mathematics they pertain to, although the dividing lines between these areas is somewhat blurry, and some articles arguably span more than one category. The articles in Sections 4.3-4.6 were written in honour of the eightieth birthday of my graduate advisor, Eli Stein, as a selection of my favourite contributions he made to analysis.

A remark on notation

For reasons of space, we will not be able to define every single mathematical term that we use in this book. If a term is italicised for reasons other than emphasis or for definition, then it denotes a standard mathematical object, result, or concept, which can be easily looked up in any number of references. (In the blog version of the book, many of these terms were linked to their Wikipedia pages, or other on-line reference pages.)

I will however mention a few notational conventions that I will use throughout. The cardinality of a finite set E will be denoted $|E|$. We will use[1] the asymptotic notation $X = O(Y)$, $X \ll Y$, or $Y \gg X$ to denote the estimate $|X| \leq CY$ for some absolute constant $C > 0$. In some cases we will need this constant C to depend on a parameter (e.g. d), in which case we shall indicate this dependence by subscripts, e.g. $X = O_d(Y)$ or $X \ll_d Y$. We also sometimes use $X \sim Y$ as a synonym for $X \ll Y \ll X$.

In many situations there will be a large parameter n that goes off to infinity. When that occurs, we also use the notation $o_{n\to\infty}(X)$ or simply $o(X)$ to denote any quantity bounded in magnitude by $c(n)X$, where $c(n)$ is a function depending only on n that goes to zero as n goes to infinity. If we need $c(n)$ to depend on another parameter, e.g. d, we indicate this by further subscripts, e.g. $o_{n\to\infty;d}(X)$.

We will occasionally use the averaging notation $\mathbf{E}_{x\in X} f(x) := \frac{1}{|X|}\sum_{x\in X} f(x)$ to denote the average value of a function $f : X \to \mathbf{C}$ on a non-empty finite set X.

If E is a subset of a domain X, we use $1_E : X \to \mathbf{R}$ to denote the *indicator function* of X, thus $1_E(x)$ equals 1 when $x \in E$ and 0 otherwise.

Acknowledgments

I am greatly indebted to many readers of my blog, Buzz, and Google+ feeds, including Andrew Bailey, Roland Bauerschmidt, Tony Carbery, Yemon Choi, Marco Frasca, Charles Gunn, Joerg Grande, Alex Iosevich, Allen Knutson, Miguel Lacruz, Srivatsan Narayanan, Andreas Seeger, Orr Shalit, David Speyer, Ming Wang, Ben Wieland, Qiaochu Yuan, Pavel Zorin, and several anonymous commenters, for corrections and other comments, which can be viewed online at

`terrytao.wordpress.com`

The author is supported by a grant from the MacArthur Foundation, by NSF grant DMS-0649473, and by the NSF Waterman award.

[1] In harmonic analysis and PDE, it is more customary to use $X \lesssim Y$ instead of $X \ll Y$.

Chapter 1

Logic and foundations

1.1. The argument from ignorance

The *argumentum ad ignorantiam* (argument from ignorance) is one of the classic fallacies in informal reasoning. In this argument, one starts with the observation that one does not know of any reason that a statement X is true (or false), and uses this as evidence to support the claim that X is therefore false (or therefore true). This argument can have a fair amount of validity in situations in which one's ability to gather information about X can be reasonably expected to be close to complete, and can give weak support for a conclusion when one's information about X is partial but substantial and unbiased (except in situations in which an adversary is deliberately exploiting gaps in this information, in which case one should proceed in a far more "game-theoretic" or "paranoid" manner). However, when dealing with statements about poorly understood phenomena, in which only a small or unrepresentative amount of data is available, the argument from ignorance can be quite dangerous, as summarised by the adage "absence of evidence is not evidence of absence".

There are versions of the "argument from ignorance" that occur in mathematics and physics; these are almost always non-rigorous arguments, but can serve as useful heuristics, or as the basis for formulating useful conjectures. Examples include the following:

(1) (Non-mathematical induction) If a statement $P(x)$ is known to be true for all computable examples of x, and one sees no reason why these examples should not be representative of the general case, then one expects $P(x)$ to be true for all x.

(2) (Principle of indifference) If a random variable X can take N different values, and there is no reason to expect one of these values to be any more likely to occur than any other, then one can expect each value to occur with probability $1/N$.

(3) (Equidistribution) If one has a (discrete or continuous) distribution of points x in a space X, and one sees no reason why this distribution should favour one portion of X over another, then one can expect this distribution to be asymptotically equidistributed in X after increasing the "sample size" of the distribution to infinity (thus, for any "reasonable" subset E of X, the portion of the distribution contained inside E should asymptotically converge to the relative measure of E inside X).

(4) (Independence) If one has two random variables X and Y, and one sees no reason why knowledge about the value of X should significantly affect the behaviour of Y (or vice versa), then one can expect X and Y to be independent (or approximately independent) as random variables.

(5) (Heuristic Borel-Cantelli) Suppose one is counting solutions to an equation such as $P(n) = 0$, where n ranges over some set N. Suppose that for any given $n \in N$, one expects the equation $P(n) = 0$ to hold with probability[1] p_n. Suppose also that one sees no significant relationship between the solvability of $P(n) = 0$ and the solvability of $P(m) = 0$ for distinct n, m. If $\sum_n p_n$ is infinite, one then expects infinitely many solutions to $P(n) = 0$; but if $\sum_n p_n$ is finite, then on expects only finitely many solutions to $P(n) = 0$.

(6) (Local-to-global principle) If one is trying to solve some sort of equation $F(x) = 0$, and all "obvious" or "local" obstructions to this solvability (e.g. trying to solve df = w when w is not closed) are not present, and one believes that the class of all possible x is so "large" or "flexible" that no global obstructions (such as those imposed by topology) are expected to intervene, then one expects a solution to exist.

The equidistribution principle is a generalisation of the principle of indifference, and among other things forms the heuristic basis for statistical mechanics (where it is sometimes referred to as the fundamental postulate of statistical mechanics). The heuristic Borel-Cantelli lemma can be viewed as a combination of the equidistribution and independence principles.

[1]Such an expectation for instance might occur from the principle of indifference, for instance by observing that $P(n)$ can range in a set of size R_n that contains zero, in which case one can predict a probability $p_n = 1/R_n$ that $P(n)$ will equal zero.

A typical example of the equidistribution principle in action is the conjecture (which is still unproven) that the digits of π are equidistributed: thus, for instance, the proportion of the first N digits of π that are equal to, say, 7, should approach $1/10$ in the limit as N goes to infinity. The point here is that we see no reason why the fractional part $\{10^n\pi\}$ of the expression $10^n\pi$ should favour one portion of the unit interval $[0,1]$ over any other, and in particular it should occupy the subinterval $[0.7, 0.8)$ one tenth of the time, asymptotically.

A typical application of the heuristic Borel-Cantelli lemma is an informal "proof" of the twin prime conjecture that there are infinitely many primes p such that $p+2$ is also prime. From the prime number theorem, we expect a typical large number n to have an (asymptotic) probability $\frac{1}{\log n}$ of being prime, and n+2 to have a probability $\frac{1}{\log(n+2)}$ of being prime. If one sees no reason why the primality (or lack thereof) of n should influence the primality (or lack thereof) of $n+2$, then by the independence principle one expects a typical number n to have a probability $\frac{1}{(\log n)(\log n+2)}$ of being the first part of a twin prime pair. Since $\sum_n \frac{1}{(\log n)(\log n+2)}$ diverges, we then expect infinitely many twin primes.

While these arguments can lead to useful heuristics and conjectures, it is important to realise that they are not remotely close to being rigorous, and can indeed lead to incorrect results. For instance, the above argument claiming to prove the infinitude of twin primes $p, p+2$ would also prove the infinitude of consecutive primes $p, p+1$, which is absurd. The reason here is that the primality of a number p *does* significantly influence the primality of its successor $p+1$, because all but one of the primes are odd, and so if p is a prime other than 2, then $p+1$ is even and cannot itself be prime. Now, this objection does not prevent $p+2$ from being prime (and neither does consideration of divisibility by 3, or 5, etc.), and so there is no obvious reason why the twin prime argument does not work; but one cannot conclude from this that there are infinitely twin primes without an appeal to the non-rigorous argument from ignorance.

Another well-known mathematical example where the argument from ignorance fails concerns the fractional parts of $\exp(\pi\sqrt{n})$, where n is a natural number. At first glance, much as with $10^n\pi$, there is no reason why these fractional parts of transcendental numbers should favour any region of the unit interval $[0,1]$ over any other, and so one expects equidistribution in n. As a consequence of this and a heuristic Borel-Cantelli argument, one expects the distance of $\exp(\pi\sqrt{n})$ to the nearest integer to not be much less than $1/n$ at best. However, as famously observed by Hermite, $\exp(\pi\sqrt{163})$ is extremely close to an integer, with the error being less than 10^{-12}. Here, there is a deeper structure present which one might previously be ignorant

of, namely the unique factorisation of the number field $\mathbf{Q}(\sqrt{-163})$. For all we know, a similar "hidden structure" or "conspiracy" might ultimately be present in the digits[2] of π, or the twin primes; we cannot yet rule these out, and so these conjectures remain open.

There are similar cautionary counterexamples that are related to the twin prime problem. The same sort of heuristics that support the twin prime conjecture also support *Schinzel's hypothesis H*, which roughly speaking asserts that polynomials $P(n)$ over the integers should take prime values for infinitely many n unless there is an "obvious" reason why this is not the case, i.e. if $P(n)$ is never coprime to a fixed modulus q, or if it is reducible, or if it cannot take arbitrarily large positive values. Thus, for instance, n^2+1 should take infinitely many prime values (an old conjecture of Landau). This conjecture is widely believed to be true, and one can use the heuristic Borel-Cantelli lemma to support it. However, it is interesting to note that if the integers $\mathbf{Z}$ are replaced by the function field analogue $\mathbf{F}_2[t]$, then the conjecture fails, as first observed by Swan [**Sw1962**]. Indeed, the octic polynomial n^8+t^3, while irreducible over $\mathbf{F}_2[t]$, turns out to never give an irreducible polynomial for any given value $n \in \mathbf{F}_2[t]$; this has to do with the structure of this polynomial in certain lifts of $\mathbf{F}_2[t]$, a phenomenon studied systematically in [**CoCoGr2008**].

Even when the naive argument from ignorance fails, though, the nature of that failure can often be quite interesting and lead to new mathematics. In my own area of research, an example of this came from the inverse theory of the Gowers uniformity norms. Naively, these norms measured the extent to which the phase of a function behaved like a polynomial, and so an argument from ignorance would suggest that the polynomial phases were the only obstructions to the Gowers uniformity norm being small; however, there was an important additional class of "pseudopolynomial phases", known as *nilsequences*, that one additionally had to consider. Proving this latter conjecture (known as the *inverse conjecture for the Gowers norms*) goes through a lot of rich mathematics, in particular the equidistribution theory of orbits in nilmanifolds, and has a number of applications, for instance in counting patterns in primes such as arithmetic progressions; see [**Ta2011b**].

1.2. On truth and accuracy

Suppose that x is an object, and X is a class of objects. What does it mean to honestly say that "x is an element of X"?

To a mathematician, the standard here is that of *truth*: the statement "x is an element of X" is honest as long as x satisfies, to the letter, absolutely

[2]Incidentally, a possible conspiracy among the digits of π is a key plot point in the novel "Contact" by Carl Sagan, though not in the more well known movie adaptation of that novel.

all of the requirements for membership in X (and similarly, "x is not an element of X" is honest if even the most minor requirement for membership is violated). Thus, for instance, a square is an example of a rectangle, a straight line segment is an example of a curve, 1 is not an example of a prime number, and so forth.

In most areas outside of mathematics, though, using strict truth as the standard for honesty is not ideal (even if people profess it to be so). To give a somewhat frivolous example, using a strict truth standard, tomatoes are not vegetables, but are technically fruits. Less frivolously, many loopholes in legal codes (such as tax codes) are based on interpretations of laws that are strictly true, but not necessarily in the spirit in which the law was intended. Even mathematicians deviate sometimes from a strict truth standard, for instance by abusing notation (e.g. using a set X when one should instead be referring to a space (such as a metric space (X, d), a measure space (X, B, μ), etc.)), or by using adverbs such as "morally" or "essentially".

In most practical situations, a better standard for honesty would be that of *accuracy* rather than *truth*. Under this standard, the statement "x is an element of X" would be honest if x is close to (or resembles) a typical element of X, with the level of honesty proportional to the degree of resemblance or closeness (and the degree of typicality). Under this standard, for instance, the assertion that a tomato is a vegetable is quite honest, as a tomato is close in practical function to a typical vegetable. On the other hand, a mathematically correct assertion such as "squares are rectangles" becomes slightly dishonest, since a generic rectangle would not have all sides equal, and so the mental image generated by labeling a square object a rectangle instead of a square is more misleading. Meanwhile, the statement "π equals $22/7$", while untrue, is reasonably accurate, and thus honest in many situations outside of higher mathematics.

Many deceptive rhetorical techniques rely on asserting statements which are true but not accurate. A good example of this is *reductio ad Hitlerum*: attacking the character of a person x by noting that x belongs to a class X which also contains Hitler. Usually, either x or Hitler (or both) will not be a typical element of X, making this attack dishonest even if all statements used in the attack are true in a strict sense. Other examples include using guilt by association, lying by omission, or by using emotionally charged words to alter the listener's perception of what a "typical" element of a class X is.

Of course, accuracy is much less of an objective standard than truth, as it is difficult to attain consensus on exactly what one means by "close" or "typical", or to decide on exactly what threshold of accuracy is acceptable for a given situation. Also, the laws of logic, which apply without exception

to truth, do not always apply without exception to accuracy. For instance, the law of the excluded middle fails: if x is a person, it is possible for the two statements "x is someone who has stopped beating his wife" and "x is someone who has not stopped beating his wife" to both[3] be dishonest. Similarly, "1 is not a prime number" and "1 is not a composite number" are true, but somewhat dishonest statements (as the former suggests that 1 is composite, while the latter suggests that 1 is prime); the joint statement "1 is neither a prime number nor a composite number" is more honest.

Ideally, of course, all statements in a given discussion should be both factually correct and accurate. But it would be a mistake to only focus on the former standard and not on the latter.

1.3. Mathematical modeling

In order to use mathematical modelling in order to solve a real-world problem, one ideally would like to have three ingredients besides the actual mathematical analysis:

(i) A good mathematical model. This is a mathematical construct which connects the observable data, the predicted outcome, and various unspecified parameters of the model to each other. In some cases, the model may be probabilistic instead of deterministic (thus the predicted outcome will be given as a random variable rather than as a fixed quantity).

(ii) A good set of observable data.

(iii) Good values for the parameters of the model.

For instance, if one wanted to work out the distance D to a distant galaxy, the model might be Hubble's law $v = HD$ relating the distance to the recessional velocity v, the data might be the recessional velocity v (or, more realistically, a proxy for that velocity, such as the red shift), and the only parameter in this case would be the Hubble constant H. This is a particularly simple situation; of course, in general one would expect a much more complex model, a much larger set of data, and a large number of parameters[4].

As mentioned above, in ideal situations one has all three ingredients: a good model, good data, and good parameters. In this case the only remaining difficulty is a direct one, namely to solve the equations of the model

[3] At the other extreme, consider Niels Bohr's quote: "The opposite of a correct statement is a false statement. But the opposite of a profound truth may well be another profound truth.".

[4] Such parameters need not be numerical; a model, for instance, could posit an unknown functional relationship between two observable quantities, in which case the function itself is the unknown parameter.

with the given data and parameters to obtain the result. This type of situation pervades undergraduate homework exercises in applied mathematics and physics, and also accurately describes many mature areas of engineering (e.g. civil engineering or mechanical engineering) in which the model, data, and parameters are all well understood. One could also classify pure mathematics as being the quintessential example of this type of situation, since the models for mathematical foundations (e.g. the ZFC model for set theory) are incredibly well understood (to the point where we rarely even think of them as models any more), and one primarily works with well-formulated problems with precise hypotheses and data.

However, there are many situations in which one or more ingredients are missing. For instance, one may have a good model and good data, but the parameters of the model are initially unknown. In that case, one needs to first solve some sort of inverse problem to recover the parameters from existing sets of data (and their outcomes), before one can then solve the direct problem. In some cases, there are clever ways to gather and use the data so that various unknown parameters largely cancel themselves out, simplifying the task. For instance, to test the efficiency of a drug, one can use a double-blind study in order to cancel out the numerous unknown parameters that affect both the control group and the experimental group equally. Typically, one cannot solve for the parameters exactly, and so one must accept an increased range of error in one's predictions. This type of problem pervades undergraduate homework exercises in statistics, and accurately describes many mature sciences, such as physics, chemistry, materials science, and some of the life sciences.

Another common situation is when one has a good model and good parameters, but an incomplete or corrupted set of data. Here, one often has to clean up the data first using error-correcting techniques before proceeding (this often requires adding a mechanism for noise or corruption into the model itself, e.g. adding gaussian white noise to the measurement model). This type of problem pervades undergraduate exercises in signal processing, and often arises in computer science and communications science.

In all of the above cases, mathematics can be utilised to great effect, though different types of mathematics are used for different situations (e.g. computational mathematics when one has a good model, data set, and parameters; statistics when one has good model and data set but unknown parameters; computer science, filtering, and compressed sensing when one has good model and parameters, but unknown data; and so forth). However, there is one important situation where the current state of mathematical sophistication is only of limited utility, and that is when it is the model which is unreliable. In this case, even having excellent data, perfect knowledge of

parameters, and flawless mathematical analysis may lead to error or a false sense of security; this for instance arose during the recent financial crisis, in which models based on independent gaussian fluctuations in various asset prices turned out to be totally incapable of describing tail events.

Nevertheless, there are still some ways in which mathematics can assist in this type of situation. For instance, one can mathematically test the robustness of a model by replacing it with other models and seeing the extent to which the results change. If it turns out that the results are largely unaffected, then this builds confidence that even a somewhat incorrect model may still yield usable and reasonably accurate results. At the other extreme, if the results turn out to be highly sensitive to the model assumptions, then even a model with a lot of theoretical justification would need to be heavily scrutinised by other means (e.g. cross-validation) before one would be confident enough to use it. Another use of mathematics in this context is to test the consistency of a model. For instance, if a model for a physical process leads to a non-physical consequence (e.g. if a partial differential equation used in the model leads to solutions that become infinite in finite time), this is evidence that the model needs to be modified or discarded before it can be used in applications.

It seems to me that one of the reasons why mathematicians working in different disciplines (e.g. mathematical physicists, mathematical biologists, mathematical signal processors, financial mathematicians, cryptologists, etc.) have difficulty communicating to each other mathematically is that their basic environment of model, data, and parameters are so different: a set of mathematical tools, principles, and intuition that works well in, say, a good model, good parameters, bad data environment may be totally inadequate or even misleading when working in, say, a bad model, bad parameters, good data environment. (And there are also other factors beyond these three that also significantly influence the mathematical environment and thus inhibit communication; for instance, problems with an active adversary, such as in cryptography or security, tend to be of a completely different nature than problems in the only adverse effects come from natural randomness, which is for instance the case in safety engineering.)

1.4. Epistemic logic, and the blue-eyed islander puzzle lower bound

In [**Ta2009**, §1.1] I discussed my favourite logic puzzle, namely the blue-eyed islander puzzle, reproduced here:

Problem 1.4.1. There is an island upon which a tribe resides. The tribe consists of 1000 people, with various eye colours. Yet, their religion forbids them to know their own eye color, or even to discuss the topic; thus, each

resident can (and does) see the eye colors of all other residents, but has no way of discovering his or her own (there are no reflective surfaces). If a tribesperson does discover his or her own eye color, then their religion compels them to commit ritual suicide at noon the following day in the village square for all to witness. All the tribespeople are highly logical[5] and devout, and they all know that each other is also highly logical and devout (and they all know that they all know that each other is highly logical and devout, and so forth).

Of the 1000 islanders, it turns out that 100 of them have blue eyes and 900 of them have brown eyes, although the islanders are not initially aware of these statistics (each of them can of course only see 999 of the 1000 tribespeople).

One day, a blue-eyed foreigner visits to the island and wins the complete trust of the tribe.

One evening, he addresses the entire tribe to thank them for their hospitality.

However, not knowing the customs, the foreigner makes the mistake of mentioning eye color in his address, remarking how unusual it is to see another blue-eyed person like myself in this region of the world.

What effect, if anything, does this *faux pas* have on the tribe?

I am fond of this puzzle because in order to properly understand the correct solution (and to properly understand why the alternative solution is incorrect), one has to think very clearly (but unintuitively) about the nature of knowledge.

There is however an additional subtlety to the puzzle that was pointed out to me, in that the correct solution to the puzzle has two components, a (necessary) upper bound and a (possible) lower bound, both of which I will discuss shortly. Only the upper bound is correctly explained in the puzzle (and even then, there are some slight inaccuracies, as will be discussed below). The lower bound, however, is substantially more difficult to establish, in part because the bound is merely possible and not necessary. Ultimately, this is because to demonstrate the upper bound, one merely has to show that a certain statement is logically deducible from an islander's state of knowledge, which can be done by presenting an appropriate chain of logical deductions. But to demonstrate the lower bound, one needs to show that certain statements are *not* logically deducible from an islander's state of knowledge, which is much harder, as one has to rule out *all* possible chains

[5]For the purposes of this logic puzzle, "highly logical" means that any conclusion that can logically deduced from the information and observations available to an islander, will automatically be known to that islander.

of deductive reasoning from arriving at this particular conclusion. In fact, to rigorously establish such impossiblity statements, one ends up having to leave the "syntactic" side of logic (deductive reasoning), and move instead to the dual "semantic" side of logic (creation of models). As we shall see, semantics requires substantially more mathematical setup than syntax, and the demonstration of the lower bound will therefore be much lengthier than that of the upper bound.

To complicate things further, the particular logic that is used in the blue-eyed islander puzzle is not the same as the logics that are commonly used in mathematics, namely *propositional logic* and *first-order logic*. Because the logical reasoning here depends so crucially on the concept of knowledge, one must work instead with an *epistemic logic* (or more precisely, an *epistemic modal logic*) which can properly work with, and model, the knowledge of various agents. To add even more complication, the role of time is also important (an islander may not know a certain fact on one day, but learn it on the next day), so one also needs to incorporate the language of *temporal logic* in order to fully model the situation. This makes both the syntax and semantics of the logic quite intricate; to see this, one only needs to contemplate the task of programming a computer with enough epistemic and temporal deductive reasoning powers that it would be able to solve the islander puzzle (or even a smaller version thereof, say with just three or four islanders) without being deliberately "fed" the solution. (The fact, therefore, that humans can actually grasp the correct solution without any formal logical training is therefore quite remarkable.)

As difficult as the syntax of temporal epistemic modal logic is, though, the semantics is more intricate still. For instance, it turns out that in order to completely model the epistemic state of a finite number of agents (such as 1000 islanders), one requires an *infinite* model, due to the existence of arbitrarily long nested chains of knowledge (e.g. "A knows that B knows that C knows that D has blue eyes"), which cannot be automatically reduced to shorter chains of knowledge. Furthermore, because each agent has only an incomplete knowledge of the world, one must take into account multiple *hypothetical worlds*, which differ from the real world but which are considered to be possible worlds by one or more agents, thus introducing *modality* into the logic. More subtly, one must also consider worlds which each agent knows to be impossible, but are not *commonly known* to be impossible, so that (for instance) one agent is willing to admit the possibility that another agent considers that world to be possible; it is the consideration of such worlds which is crucial to the resolution of the blue-eyed islander puzzle. And this is even before one adds the temporal aspect (e.g. "On Tuesday, A knows that on Monday, B knew that by Wednesday, C will know that D has blue eyes").

Despite all this fearsome complexity, it *is* still possible to set up both the syntax and semantics of temporal epistemic modal logic[6] in such a way that one can formulate the blue-eyed islander problem rigorously, and in such a way that one has both an upper and a lower bound in the solution. The purpose of this section is to construct such a setup and to explain the lower bound in particular. The same logic is also useful for analysing another well-known paradox, the *unexpected hanging paradox*, and I will do so at the end of this section. Note though that there is more than one way[7] to set up epistemic logics, and they are not all equivalent to each other.

Our approach here will be a little different from the approach commonly found in the epistemic logic literature, in which one jumps straight to "arbitrary-order epistemic logic" in which arbitrarily long nested chains of knowledge ("A knows that B knows that C knows that ...") are allowed. Instead, we will adopt a hierarchical approach, recursively defining for $k = 0, 1, 2, \ldots$ a "k^{th}-order epistemic logic" in which knowledge chains of depth up to k, but no greater, are permitted. The arbitrarily order epistemic logic is then obtained as a limit (a *direct limit* on the syntactic side, and an *inverse limit* on the semantic side, which is dual to the syntactic side) of the finite order epistemic logics. The relationship between the traditional approach (allowing arbitrarily depth from the start) and the hierarchical one presented here is somewhat analogous to the distinction between *Zermelo-Fraenkel-Choice* (ZFC) set theory without the *axiom of foundation*, and ZFC with that axiom.

I should warn that this is going to be a rather formal and mathematical article. Readers who simply want to know the answer to the islander puzzle would probably be better off reading the discussion at

`terrytao.wordpress.com/2011/04/07/the-blue-eyed-islanders-puzzle-repost`

.

I am indebted to Joe Halpern for comments and corrections.

1.4.1. Zeroth-order logic. Before we plunge into the full complexity of epistemic logic (or temporal epistemic logic), let us first discuss formal logic in general, and then focus on a particularly simple example of a logic, namely *zeroth order logic* (better known as *propositional logic*). This logic will end up forming the foundation for a hierarchy of epistemic logics, which will be needed to model such logic puzzles as the blue-eyed islander puzzle.

[6] On the other hand, for puzzles such as the islander puzzle in which there are only a finite number of atomic propositions and no free variables, one at least can avoid the need to admit *predicate logic*, in which one has to discuss quantifiers such as $\forall$ and $\exists$. A fully formed predicate temporal epistemic modal logic would indeed be of terrifying complexity.

[7] In particular, one can also proceed using *Kripke models* for the semantics, which in my view, are more elegant, but harder to motivate than the more recursively founded models presented here.

Informally, a logic consists of three inter-related components:

(1) A *language*. This describes the type of sentences the logic is able to discuss.

(2) A *syntax* (or more precisely, a *formal system* for the given language). This describes the rules by which the logic can deduce conclusions (from given hypotheses).

(3) A *semantics*. This describes the sentences which the logic interprets to be true (in given models).

A little more formally:

(1) A *language* is a set L of *sentences*, which are certain strings of symbols from a fixed alphabet, that are generated by some rules of grammar.

(2) A *syntax* is a collection of *inference rules* for generating deductions of the form $T \vdash S$ (which we read as "From T, we can deduce S" or "S is a consequence of T"), where T and S are sentences in L (or sets of sentences in L).

(3) A *semantics* describes what a *model* (or *interpretation*, or *structure*, or *world*) M of the logic is, and defines what it means for a sentence S in L (or a collection of sentences) to be true in such a model M (which we write as $M \models S$, and we read as "M models S", "M obeys S", or "S is true in M").

We will abuse notation a little bit and use the language L as a metonym for the entire logic; strictly speaking, the logic should be a tuple $(L, \vdash_L, \models_L)$ consisting of the language, syntax, and semantics, but this leads to very unwieldy notation.

The syntax and semantics are dual to each other in many ways; for instance, the syntax of deduction can be used to show that certain statements can be proved, while the semantics can be used to show that certain statements *cannot* be proved. This distinction will be particularly important in the blue-eyed islander puzzle; in order to show that all blue-eyed islanders commit suicide by the 100^{th} day, one can argue purely on formal syntactical grounds; but to show that it is possible for the blue-eyed islanders to *not* commit suicide on the 99^{th} day or any preceding day, one must instead use semantic methods.

To illustrate the interplay between language, deductive syntax, and semantics, we begin with the simple example of *propositional logic*. To describe this logic, one must first begin with some collection of *atomic propositions*. For instance, on an island with three islanders I_1, I_2, I_3, one could consider

the propositional logic generated by three atomic propositions A_1, A_2, A_3, where each A_i is intended to model the statement that I_i has blue eyes.

One can have either a finite or an infinite set of atomic propositions. In this discussion, it will suffice to consider the situation in which there are only finitely many atomic propositions, but one can certainly also study logics with infinitely many such propositions.

The language L would then consist of all the sentences that can be formed from the atomic propositions using the usual logical connectives ($\wedge$, $\vee$, $\neg$, $\implies$, $\top$, $\bot$, etc.) and parentheses, according to the usual rules of logical grammar (which consists of rules such as "If S and T are sentences in L, then $(S \vee T)$ is also a sentence in L"). For instance, if A_1, A_2, A_3 are atomic propositions, then

$$((A_1 \wedge A_2) \vee (A_3 \wedge \neg A_1))$$

would be an example of a sentence in L. On the other hand,

$$\wedge A_1 \neg A_3 A_1) \vee \implies A_2($$

is not a sentence in L, despite being a juxtaposition of atomic propositions, connectives, and parentheses, because it is not built up from rules of grammar.

One could certainly write down a finite list of all the rules of grammar for propositional calculus (as is done in any basic textbook on mathematical logic), but we will not do so here in order not to disrupt the flow of discussion.

It is customary to abuse notation slightly and omit parentheses when they are redundant (or when there is enough associativity present that the precise placement of parentheses are not relevant). For instance, $((A_1 \wedge A_2) \wedge A_3)$ could be abbreviated as $A_1 \wedge A_2 \wedge A_3$. We will adopt this type of convention in order to keep the exposition as uncluttered as possible.

Now we turn to the syntax of propositional logic. This syntax is generated by basic rules of deductive logic, such as *modus ponens*

$$A, (A \implies B) \vdash B$$

or the *law of the excluded middle*

$$\vdash (A \vee \neg A)$$

and completed by transitivity (if $S \vdash T$ and $T \vdash U$, then $S \vdash U$), monotonicity ($S, T \vdash S$), and concatenation (if $S \vdash T$ and $S \vdash U$ then $S \vdash T, U$). (Here we adopt the usual convention of representing a set of sentences without using the usual curly braces, instead relying purely on the comma separator.) Another convenient inference rule to place in this logic is the *deduction theorem*: if $S \vdash T$, then one can infer $\vdash (S \implies T)$. In propositional logic (or predicate logic), this rule is redundant (hence the designation of this rule as

a *theorem*), but for the epistemic logics below, it will be convenient to make deduction an explicit inference rule, as it simplifies the other inference rules one will have to add to the system.

A typical deduction that comes from this syntax is

$$(A_1 \vee A_2 \vee A_3), \neg A_2, \neg A_3 \vdash A_1$$

which using the blue-eyed islander interpretation, is the formalisation of the assertion that given that at least one of the islanders I_1, I_2, I_3 has blue eyes, and that I_2, I_3 do not have blue eyes, one can deduce that I_1 has blue eyes.

As with the laws of grammar, one can certainly write down a finite list of inference rules in propositional calculus; again, such lists may be found in any text on mathematical logic. Note though that, much as a given vector space has more than one set of generators, there is more than one possible list of inference rules for propositional calculus, due to some rules being equivalent to, or at least deducible from, other rules; the precise choice of basic inference rules is to some extent a matter of personal taste and will not be terribly relevant for the current discussion.

Finally, we discuss the semantics of propositional logic. For this particular logic, the models M are described by *truth assignments*, that assign a truth value $(M \models A_i) \in \{\text{true}, \text{false}\}$ to each atomic statement A_i. Once a truth value $(M \models A_i)$ to each atomic statement A_i is assigned, the truth value $(M \models S)$ of any other sentence S in the propositional logic generated by these atomic statements can then be interpreted using the usual *truth tables*. For instance, returning to the islander example, consider a model M in which $M \models A_1$ is true, but $M \models A_2$ and $M \models A_3$ are false; informally, M describes a hypothetical world in which I_1 has blue eyes but I_2 and I_3 do not have blue eyes. Then the sentence $A_1 \vee A_2 \vee A_3$ is true in M,

$$M \models (A_1 \vee A_2 \vee A_3),$$

but the statement $A_1 \implies A_2$ is false in M,

$$M \not\models (A_1 \implies A_2).$$

If S is a set of sentences, we say that M models S if M models each sentence in S. Thus for instance, if we continue the preceding example, then

$$M \models (A_1 \vee A_2 \vee A_3), (A_2 \implies A_3)$$

but

$$M \not\models (A_1 \vee A_2 \vee A_3), (A_1 \implies A_2).$$

Note that if there are only finitely many atomic statements $A_1, \ldots, A_n$, then there are only finitely many distinct models M of the resulting propositional logic; in fact, there are exactly 2^n such models, one for each truth

assignment. We will denote the space of all possible models of a language L as $\mathrm{Mod}(L)$.

If one likes, one can designate one of these models to be the "real" world Real so that all the other models become purely hypothetical worlds. In the setting of propositional logic, the hypothetical worlds then have no direct bearing on the real world; the fact that a sentence S is true or false in a hypothetical world M does not say anything about what sentences are true or false in Real. However, when we turn to epistemic logics later in this section, we will see that hypothetical worlds will play an important role in the real world, because such worlds may be considered to be possible worlds by one or more agents (or, an agent may consider it possible that another agent considers the world to be possible, and so forth.).

The syntatical and semantic sides of propositional logic are tied together by two fundamental facts:

Theorem 1.4.2 (Soundness and completeness)**.** *Let L be a propositional logic, and let S be a set of sentences in L, and let T be another sentence in L.*

(1) *(Soundness) If $S \vdash T$, then every model M which obeys S, also obeys T (i.e. $M \models S$ implies $M \models T$).*

(2) *(Completeness) If every model M that obeys S, also obeys T, then $S \vdash T$.*

Soundness is easy to prove; one merely needs to verify that each of the inference rules $S \vdash T$ in one's syntax is *valid*, in that models that obey S, automatically obey T. This boils down to some tedious inspection of truth tables. (The soundness of the deduction theorem is a little trickier to prove, but one can achieve this by an induction on the number of times this theorem is invoked in a given induction.) Completeness is a bit more difficult to establish; this claim is in fact a special case of the *Gödel completeness theorem*, and is discussed in [**Ta2010b**, §1.4]; we also sketch a proof of completeness below.

By taking the contrapositive of soundness, we have the following important corollary: if we can find a model M which obeys S but does not obey T, then it must not be possible to deduce T as a logical consequence of S: $S \nvdash T$. Thus, *we can use semantics to demonstrate limitations in syntax.*

For instance, consider a truth assignment M in which A_2 is true but A_1 is false. Then $M \models (A_1 \implies A_2)$, but $M \not\models (A_2 \implies A_1)$. This demonstrates that

$$(A_1 \implies A_2) \nvdash (A_2 \implies A_1),$$

thus an implication such as $A_1 \implies A_2$ does not entail its converse $A_2 \implies A_1$.

A *theory* (or more precisely, a *deductive theory*) in a logic L, is a set of sentences $\mathcal{T}$ in L which is closed under deductive consequence, thus if $\mathcal{T} \vdash S$ for some sentence S in L, then $S \in \mathcal{T}$. Given a theory $\mathcal{T}$, one can associate the set

$$\mathrm{Mod}_L(\mathcal{T}) = \mathrm{Mod}(\mathcal{T}) := \{M \in \mathrm{Mod}(L) : M \models \mathcal{T}\}$$

of all possible worlds (or models) in which that theory is true; conversely, given a set $\mathcal{M} \subset \mathrm{Mod}(L)$ of such models, one can form the *theory*

$$\mathrm{Th}_L(\mathcal{M}) = \mathrm{Th}(\mathcal{M}) := \{S \in L : M \models S \text{ for all } M \in \mathcal{M}\}$$

of sentences which are true in all models in $\mathcal{M}$. If the logic L is both sound and complete, these operations invert each other: given any theory $\mathcal{T}$, we have $\mathrm{Th}(\mathrm{Mod}(\mathcal{T})) = \mathcal{T}$, and given any set $\mathcal{M} \subset \mathrm{Mod}(L)$ of models, $\mathrm{Th}(\mathcal{M})$ is a theory and $\mathrm{Mod}(\mathrm{Th}(\mathcal{M})) = \mathcal{M}$. Thus there is a one-to-one correspondence between theories and sets of possible worlds in a sound complete language L.

For instance, in our running example, if $\mathcal{T}$ is the theory generated by the three statements $A_1 \vee A_2 \vee A_3$, A_2, and $\neg A_3$, then $\mathrm{Mod}(\mathcal{T})$ consists of precisely two worlds; one in which A_1, A_2 are true and A_3 is false, and one in which A_2 is true and A_1, A_3 are false. Since neither of A_1 or $\neg A_1$ are true in both worlds in $\mathrm{Mod}(\mathcal{T})$, neither of A_1 or $\neg A_1$ lie in $\mathcal{T} = \mathrm{Th}(\mathrm{Mod}(\mathcal{T}))$. Thus, it is not possible to deduce either of A_1 or $\neg A_1$ from the hypotheses $A_1 \vee A_2 \vee A_3$, A_2, and $\neg A_3$. More informally, if one knows that there is at least one blue-eyed islander, that I_2 has blue eyes, and I_3 does not have blue eyes, this is not enough information to determine whether I_1 has blue eyes or not.

One can use theories to prove the completeness theorem. Roughly speaking, one can argue by taking the contrapositive. Suppose that $S \nvdash T$, then we can find a theory which contains all sentences in S, but does not contain T. In this finite setting, we can easily pass to a maximal such theory (with respect to set inclusion); one then easily verifies that this theory is *complete* in the sense that for any given sentence U, exactly one of U and $\neg U$ is true. From this complete theory one can then directly build a model M which obeys S but does not obey T, giving the desired claim.

1.4.2. First-order epistemic logic. Having reviewed propositional logic (which we will view as the zeroth-order iteration of epistemic logic), we now turn to the first non-trivial example of epistemic logic, which we shall call *first-order epistemic logic* (which should not be confused with the more familiar *first-order predicate logic*). Roughly speaking, first-order epistemic logic is like zeroth-order logic, except that there are now also some knowledge

agents that are able to know certain facts in zeroth-order logic (e.g. an islander I_1 may know that the islander I_2 has blue eyes). However, in this logic one cannot yet express higher-order facts (e.g. we will not yet be able to formulate a sentence to the effect that I_1 knows that I_2 knows that I_3 has blue eyes). This will require a second-order or higher epistemic logic, which we will discuss later in this section.

Let us now formally construct this logic. As with zeroth-order logic, we will need a certain set of atomic propositions, which for simplicity we will assume to be a finite set $A_1, \ldots, A_n$. This already gives the zeroth order language L_0 of sentences that one can form from the $A_1, \ldots, A_n$ by the rules of propositional grammar. For instance,

$$(A_1 \implies A_2) \wedge (A_2 \implies A_3)$$

is a sentence in L_0. The zeroth-order logic L_0 also comes with a notion of inference $\vdash_{L_0}$ and a notion of modeling $\models_{L_0}$, which we now subscript by L_0 in order to distinguish it from the first-order notions of inference $\vdash_{L_1}$ and modeling $\models_{L_1}$ which we will define shortly. Thus, for instance

$$(A_1 \implies A_2) \wedge (A_2 \implies A_3) \vdash_{L_0} (A_1 \implies A_3),$$

and if M_0 is a truth assignment for L_0 for which A_1, A_2, A_3 are all true, then

$$M_0 \models_{L_0} (A_1 \implies A_2) \wedge (A_2 \implies A_3).$$

We will also assume the existence of a finite number of *knowledge agents* $K_1, \ldots, K_m$, each of which are capable of knowing sentences in the zeroth order language L_0. (In the case of the islander puzzle, and ignoring for now the time aspect of the puzzle, each islander I_i generates one knowledge agent K_i, representing the state of knowledge of I_i at a fixed point in time. Later on, when we add in the temporal aspect to the puzzle, we will need different knowledge agents for a single islander at different points in time, but let us ignore this issue for now.) To formalise this, we define the first-order language L_1 to be the language generated from L_0 and the rules of propositional grammar by imposing one additional rule:

- If S is a sentence in L_0, and K is a knowledge agent, then $K(S)$ is a sentence in L_1 (which can informally be read as "K knows (or believes) S to be true").

Thus, for instance,

$$K_2(A_1) \wedge K_1(A_1 \vee A_2 \vee A_3) \wedge \neg A_3$$

is a sentence in L_1; in the islander interpretation, this sentence denotes the assertion that I_2 knows I_1 to have blue eyes, and I_1 knows that at least one islander has blue eyes, but I_3 does not have blue eyes. On the other hand,

$$K_1(K_2(A_3))$$

is *not* a sentence in L_1, because $K_2(A_3)$ is not a sentence in L_0. (However, we will be able to interpret $K_1(K_2(A_3))$ in the second-order epistemic language L_2 that we will define later.)

We give L_1 all the rules of syntax that L_0 presently enjoys. For instance, thanks to modus ponens, we have

$$K_1(A_1) \wedge (K_1(A_1) \implies K_1(A_2)) \vdash_{L_1} K_1(A_2). \tag{1.1}$$

Similarly, if S, T are sentences in L_0 such that $S \vdash_{L_0} T$, then one automatically has $S \vdash_{L_1} T$.

However, we would like to add some additional inference rules to reflect our understanding of what "knowledge" means. One has some choice in deciding what rules to lay down here, but we will only add one rule, which informally reflects the assertion that "all knowledge agents are highly logical":

- **First-order epistemic inference rule:** If $S_1, \ldots, S_i, T \in L_0$ are sentences such that
$$S_1, \ldots, S_i \vdash_{L_0} T$$
and K is a knowledge agent, then
$$K(S_1), \ldots, K(S_i) \vdash_{L_1} K(T).$$

We will introduce higher order epistemic inference rules when we turn to higher order epistemic logics.

Informally speaking, the epistemic inference rule asserts that if T can be deduced from $S_1, \ldots, S_i$, and K knows $S_1, \ldots S_i$ to be true, then K must also know T to be true. For instance, since *modus ponens* gives us the inference

$$A_1, (A_1 \implies A_2) \vdash_{L_0} A_2$$

we therefore have, by the first-order epistemic inference rule,

$$K_1(A_1), K_1(A_1 \implies A_2) \vdash_{L_1} K_1(A_2)$$

(note how this is different from (1.1) - why?).

Another example of more relevance to the islander puzzle, we have

$$(A_1 \vee A_2 \vee A_3), \neg A_2, \neg A_3 \vdash_{L_0} A_1$$

and thus, by the first-order epistemic inference rule,

$$K_1(A_1 \vee A_2 \vee A_3), K_1(\neg A_2), K_1(\neg A_3) \vdash_{L_1} K_1(A_1).$$

In the islander interpretation, this asserts that if I_1 knows that one of the three islanders I_1, I_2, I_3 has blue eyes, but also knows that I_2 and I_3 do not have blue eyes, then I_1 must also know that he himself (or she herself) has blue eyes.

One particular consequence of the first-order epistemic inference rule is that if a sentence $T \in L_0$ is a *tautology* in L_0 - true in every model of L_0, or equivalently (by completeness) deducible from the inference rules of L_0, and K is a knowledge agent, then $K(T)$ is a tautology in L_1: $\vdash_{L_0} T$ implies $\vdash_{L_1} K(T)$. Thus, for instance, we have $\vdash_{L_1} K_1(A_1 \implies A_1)$, because A_1 is a tautology in L_0 (thus $\vdash_{L_0} A_1 \implies A_1$).

It is important to note, however, that if a statement T is not a tautology, but merely *true* in the "real" world Real, this does *not* imply that $K(T)$ is also true in the real world: as we shall see later, Real $\models_{L_0} T$ does not imply Real $\models_{L_1} K(T)$. (We will define what $\models_{L_1}$ means presently.) Intuitively, this reflects the obvious fact that knowledge agents need not be omniscient; it is possible for a sentence T to be true without a given agent K being aware of this truth.

In the converse direction, we also allow for the possibility that $K(T)$ is true in the real world, without T being true in the real world, thus it is conceivable that Real $\models_{L_1} K(T)$ is true but Real $\models_{L_0} T$ is false. This reflects the fact that a knowledge agent may in fact have incorrect knowledge of the real world. (This turns out not to be an important issue in the islander puzzle, but is of relevance for the unexpected hanging puzzle.)

In a related spirit, we also allow for the possibility that $K(T)$ and $K(\neg T)$ may both be true in the real world; an agent may conceivably be able to know inconsistent facts. However, from the inference $T, \neg T \vdash_{L_0} S$ of *ex falso quodlibet* and the first-order epistemic inference rule, this would mean that $K(S)$ is true in this world for every S in L_0, thus this knowledge agent believes absolutely every statement to be true. Again, such inconsistencies are not of major relevance to the islander puzzle, but as we shall see, their analysis is important for resolving the unexpected hanging puzzle correctly.

Remark 1.4.3. It is perhaps worth re-emphasising the previous points. In some interpretations of knowledge, $K(S)$ means that S has somehow been "justified" to be true, and in particular $K(S)$ should entail S in such interpretations. However, we are taking a more general (and abstract) point of view, in which we are agnostic as regards to whether K represents necessary or justified knowledge. In particular, our analysis also applies to "generalised knowledge" operators, such as "belief". One can of course specialise this general framework to a more specific knowledge concept by adding more axioms, in which case one can obtain sharper conclusions regarding the resolution of various paradoxes, but we will work here primarily in the general setting.

Having discussed the language and syntax of the first-order epistemic logic L_1, we now turn to the semantics, in which we describe the possible

models M_1 of L_1. As L_1 is an extension of L_0, any model M_1 of L_1 must contain as a component a model M_0 of L_0, which describes the truth assignment of each of the atomic propositions A_i of L_0; but it must also describe the state of knowledge of each of the agents K_i in this logic. One can describe this state in two equivalent ways; either as a *theory* $\{S \in L_0 : M_1 \models_{L_1} K_i(S)\}$ (in L_0) of all the sentences S in L_0 that K_i knows to be true (which, by the first-order epistemic inference rule, is closed under $\vdash_{L_0}$ and is thus indeed a theory in L_0); or equivalently (by the soundness and completeness of L_0), as a set

$$\{M_{0,i} \in \mathrm{Mod}(L_0) : M_{0,i} \models_{L_0} S \text{ whenever } M_1 \models_{L_1} K_i(S)\}$$

of all the possible models of L_0 in which all the statements that K_i knows to be true, are in fact true. We will adopt the latter perspective; thus a model M_1 of L_1 consists of a tuple

$$M_1 = (M_0, \mathcal{M}_0^{(1)}, \ldots, \mathcal{M}_0^{(m)})$$

where $M_0 \in \mathrm{Mod}(L_0)$ is a model of L_0, and for each $i = 1, \ldots, m$, $\mathcal{M}_0^{(i)} \subset \mathrm{Mod}(L_0)$ is a set of models of L_0. To interpret sentences $S \in L_1$ in M_1, we then declare $M_1 \models A_i$ iff $M_0 \models A_i$ for each atomic sentence A_i, and declare $M_1 \models K_i(S)$ iff S is true in every model in $\mathcal{M}_0^{(i)}$, for each $i = 1, \ldots, m$ and $S \in L_0$. All other sentences in L_1 are then interpreted by applying the usual truth tables.

As an example of such a model, consider a world with three islanders I_1, I_2, I_3, each of which has blue eyes, and can each see that each other has blue eyes, but are each unaware of their own eye colour. In this model, M_0 assigns a true value to each of A_1, A_2, A_3. As for $\mathcal{M}_0^{(1)}$, which describes the knowledge state of I_1, this set consists of two possible L_0-worlds. One is the "true" L_0-world M_0, in which A_1, A_2, A_3 are all true; but there is also an additional hypothetical L_0-world $M_{0,1}$, in which A_2, A_3 is true but A_1 is false. With I_1's current state of knowledge, neither of these two possibilities can be ruled out. Similarly, $\mathcal{M}_0^{(2)}$ and $\mathcal{M}_0^{(3)}$ will also consist of two L_0-worlds, one of which is the "true" L_0-world M_0, and the other is not.

In this particular case, the true L_0-world M_0 is included as a possible world in each of the knowledge agents' set of possible worlds $\mathcal{M}_0^{(i)}$, but in situations in which the agents knowledge is incorrect or inconsistent, it can be possible for M_0 to not be an element of one or more of the $\mathcal{M}_0^{(i)}$.

Remark 1.4.4. One can view an L_1 model M_1 as consisting of the "real world" - the L_0-model M_0 - together with m clouds $\mathcal{M}_0^{(i)}$, $i = 1, \ldots, m$ of "hypothetical worlds", one for each knowledge agent K_i. If one chooses, one can "enter the head" of any one of these knowledge agents K_i to see what he or she is thinking. One can then select any one of the L_0-worlds $M_{0,i}$

in $\mathcal{M}_0^{(i)}$ as a "possible world" in K_i's worldview, and explore that world further. Later on we will iterate this process, giving a tree-like structure to the higher order epistemic models.

Let $\mathrm{Mod}(L_1)$ be the set of all models of L_1. This is quite a large set; if there are n atomic statements $A_1, \ldots, A_n$ and m knowledge agents $K_1, \ldots, K_m$, then there are 2^n possibilities for the L_0-world M_0, and each knowledge agent K_i has its own independent set $\mathcal{M}_0^{(i)}$ of possible worlds, of which there are 2^{2^n} different possibilities, leading to 2^{n+m2^n} distinct models M_1 for L_1 in all. For instance, with three islanders wondering about eye colours, this leads to 2^{27} possibilities (although, once everyone learns each other's eye colour, the number of possible models goes down quite significantly).

It can be shown (but is somewhat tedious to do so) that the syntax and semantics of the first-order epistemic logic L_1 is still sound and complete, basically by mimicking (and using) the proof of the soundness and completeness of L_0; we sketch a proof of this below when we discuss higher order logics.

1.5. Higher-order epistemic logic

We can iterate the above procedure and construct a language, syntax, and semantics for k^{th} order epistemic logic L_k generated by some atomic propositions $A_1, \ldots, A_n$ and knowledge agents $K_1, \ldots, K_m$, recursively in terms of the preceding epistemic logic L_{k-1}. More precisely, let $k \geq 1$ be a natural number, and suppose that the logic L_{k-1} has already been defined. We then define the language of L_k as the extension of L_{k-1} generated by the laws of propositional grammar and the following rule:

- If S is a sentence in L_{k-1}, and K is a knowledge agent, then $K(S)$ is a sentence in L_k.

Thus, for instance, in the running example of three propositions A_1, A_2, A_3 and three knowledge agents K_1, K_2, K_3,

$$K_1(A_3) \wedge K_1(K_2(A_3))$$

is a sentence in L_2 (and hence in L_3, L_4, etc.) but not in L_1.

As for the syntax, we adopt all the inference rules of ordinary propositional logic, together with one new rule:

- **k^{th}-order epistemic inference rule:** If $S_1, \ldots, S_i, T \in L_{k-1}$ are sentences such that

$$S_1, \ldots, S_i \vdash_{L_{k-1}} T$$

and K is a knowledge agent, then

$$K(S_1), \ldots, K(S_i) \vdash_{L_k} K(T).$$

Thus, for instance, starting with

$$A_1, (A_1 \implies A_2) \vdash_{L_0} A_2$$

one has

$$K_1(A_1), K_1(A_1 \implies A_2) \vdash_{L_1} K_1(A_2),$$

and then

$$K_2(K_1(A_1)), K_2(K_1(A_1 \implies A_2)) \vdash_{L_2} K_2(K_1(A_2)),$$

and so forth. Informally, this rule asserts that all agents are highly logical, that they know that all agents are highly logical, and so forth. A typical deduction from these inference rules, which is again of relevance to the islander puzzle, is

$$K_1(K_2(A_1 \vee A_2 \vee A_3)), K_1(K_2(\neg A_3)) \vdash_{L_2}$$
$$K_1((\neg K_2(A_2)) \implies (\neg K_2(\neg A_1))).$$

Remark 1.5.1. This is a very minimal epistemic syntax, and is weaker than some epistemic logics considered in the literature. For instance, we do not have any version of the *positive introspection rule*

$$K(S) \vdash K(K(S));$$

thus we allow the possibility that an agent knows S "subconsciously", in that the agent knows S but does not know that he or she knows S. Similarly, we do not have any version of the *negative introspection rule*

$$\neg K(S) \vdash K(\neg K(S)),$$

so we allow the possibility that an agent is "unaware of his or her own ignorance". One can of course add these additional rules *ex post facto* and see how this strengthens the syntax and limits the semantics, but we will not need to do so here.

There is also no reason to expect the knowledge operators to commute:

$$K(K'(S)) \nvdash K'(K(S)).$$

Now we turn to the semantics. A model M_k of the language L_k consists of a L_0-model $M_0 \in \mathrm{Mod}(L_0)$, together with sets of possible L_{k-1}-models $\mathcal{M}_{k-1}^{(1)}, \ldots, \mathcal{M}_{k-1}^{(m)} \subset \mathrm{Mod}(L_{k-1})$ associated to their respective knowledge agents $K_1, \ldots, K_m$. To describe how M_k models sentences, we declare $M_k \models_{L_k} A_i$ iff $M_0 \models_{L_0} A_i$, and for any sentence S in L_{k-1} and $i = 1, \ldots, m$, we declare $M_k \models_{L_k} K_i(S)$ iff one has $M_{k-1,i} \models S$ for every $M_{k-1} \in \mathcal{M}_{k-1}^{(i)}$.

Example 1.5.2. We consider an islander model with n atomic propositions $A_1, \ldots, A_n$ (with each A_i representing the claim that I_i has blue eyes) and n knowledge agents $K_1, \ldots, K_n$ (with K_i representing the knowledge state of I_i at a fixed point in time). There are 2^n L_0-models M_0, determined by the truth values they assign to the n atomic propositions $A_1, \ldots, A_n$. For each $k \geq 0$, we can then recursively associate a L_k-model $M_k(M_0)$ to each L_0-model M_0, by setting $M_0(M_0) := M_0$, and then for $k \geq 1$, setting $M_k(M_0)$ to be the L_k-model with L_0-model M_0, and with $\mathcal{M}_{k-1}^{(i)}$ consisting of the pair $\{M_{k-1}(M_{0,i}^-), M_{k-1}(M_{0,i}^+)\}$, where $M_{0,i}^-$ (resp. $M_{0,i}^+$) is the L_0-model which is identical to M_0 except that the truth value of A_i is set to false (resp. true). Informally, $M_k(M_0)$ models the k^{th}-order epistemology of the L_0-world M_0, in which each islander sees each other's eye colour (and knows that each other islander can see all other islander's eye colour, and so forth for k iterations), but is unsure as to his or her own eye colour (which is why the set $\mathcal{M}_{k-1}^{(i)}$ of A_i's possible L_{k-1}-worlds branches into two possibilities). As one recursively explores the clouds of hypothetical worlds in these models, one can move further and further away from the "real" world. Consider for instance the situation when $n = 3$ and $M_0 \models A_1, A_2, A_3$ (thus in the "real" world, all three islanders have blue eyes), and $k = 3$. From the perspective of K_1, it is possible that one is in the world $M_2(M_{0,1}^-)$, in which I_1 does not have blue eyes: $M_{0,1}^- \models \neg A_1, A_2, A_3$. In that world, we can then pass to the perspective of K_2, and then one could be in the world $M_1(M_{0,1,2}^{--})$, in which neither I_1 nor I_2 have blue eyes: $M_{0,1,2}^{--} \models \neg A_1, \neg A_2, A_3$. Finally, inside this doubly nested hypothetical world, one can consider the perspective of K_3, in which one could be in the world $M_{0,1,2,3}^{---}$, in which none of I_1, I_2, I_3 have blue eyes: $M_{0,1,2,3}^{---} \models \neg A_1, \neg A_2, \neg A_3$. This is the total opposite of the "real" model M_0, but cannot be ruled out in at this triply nested level. In particular, we have

$$M_3(M_0) \models \neg K_1(K_2(K_3(A_1 \vee A_2 \vee A_3)))$$

despite the fact that

$$M_3(M_0) \models A_1 \vee A_2 \vee A_3$$

and

$$M_3(M_0) \models K_i(A_1 \vee A_2 \vee A_3)$$

and

$$M_3(M_0) \models K_i(K_j(A_1 \vee A_2 \vee A_3))$$

for all $i, j \in \{1, 2, 3\}$. (In particular, the statement $A_1 \vee A_2 \vee A_3$, which asserts "at least one islander has blue eyes", is not *common knowledge* in $M_3(M_0)$.

We have the basic soundness and completeness properties:

Proposition 1.5.3. *For each $k \geq 0$, L_k is both sound and complete.*

Proof. (Sketch) This is done by induction on k. For $k = 0$, this is just the soundness and completeness of propositional logic. Now suppose inductively that $k \geq 1$ and the claim has already been proven for $k-1$. Soundness can be verified as in the propositional logic case (with the validity of the k^{th} epistemic inference rule being justified by induction). For completeness, one again uses the trick of passing to a maximal L_k-theory $\mathcal{T}$ that contains one set S of sentences in L_k, but not another sentence T. This maximal L_k-theory $\mathcal{T}$ uniquely determines an L_0-model M_0 by inspecting whether each A_i or its negation lies in the theory, and also determines L_{k-1}-theories $\{S \in L_{k-1} : K_i(S) \in \mathcal{T}\}$ for each $i = 1, \ldots, m$. By induction hypothesis, each of these theories can be identified with a collection $\mathcal{M}_{k-1}^{(i)}$ of L_{k-1}-models, thus creating a L_k-model M_k that obeys T but not S, giving (the contrapositive of) completeness. □

1.5.1. Arbitrary order epistemic logic. An easy induction shows that the k^{th} order logic L_k extends the previous logic L_{k-1}, in the sense that every sentence in L_{k-1} is a sentence in L_k, every deduction on L_{k-1} is also a deduction in L_k, and every model of L_k projects down (by "forgetting" some aspects of the model) to a model of L_{k-1}. We can then form a limiting logic L_ω, whose language is the union of all the L_k (thus, S is a sentence in L_ω iff S lies in L_k for some k), whose deductive implications are the union of all the L_k deductive implications (thus, $S \vdash_{L_\infty} T$ if we have $(S \cap L_k) \vdash_{L_k} T$ for some k), and whose models are the *inverse limits* of the L_k models (thus, a model M_∞ of L_∞ is an infinite sequence of models M_k of L_k for each k, such that each M_k projects down to M_{k-1} for $k \geq 1$. It is not difficult to see that the soundness and completeness of each of the L_k implies the soundness and completeness of the limit L_∞ (assuming the axiom of choice, of course, in our metamathematics).

Remark 1.5.4. These models M_∞ are not quite the usual models of L_∞ one sees in the literature, namely *Kripke models*; roughly speaking, the models here are those Kripke models which are "well-founded" in some sense, in that they emerge from a hierarchical construction. Conversely, a Kripke model in our notation would be a collection $\mathcal{W}$ of worlds, with each world W_∞ in $\mathcal{W}$ associated with an L_0-model W_0, as well as sets $M_{K,\infty}(W_\infty) \subset \mathcal{W}$ for each knowledge agent K, describing all the worlds in $\mathcal{W}$ that K considers possible in W. Such models can be shown to be identifiable (in the sense that they give equivalent semantics) with the models described hierarchically as the inverse limits of finite depth models, but we will not detail this here.

The logic L_∞ now allows one to talk about arbitrarily deeply nested strings of knowledge: if S is a sentence in L_∞, and K is a knowledge agent, then $K(S)$ is also a sentence in L_∞. This allows for the following definition:

Definition 1.5.5 (Common knowledge)**.** If S is a sentence in L_∞, then $C(S)$ is the set of all sentences of the form

$$K_{i_1}(K_{i_2}(\ldots(K_{i_k}(S))\ldots))$$

where $k \geq 0$ and $K_{i_1}, \ldots, K_{i_k}$ are knowledge agents (possibly with repetition).

Thus, for instance, using the epistemic inference rules, every tautology in L_∞ is commonly known as such: if $\vdash_{L_\infty} S$, then $\vdash_{L_\infty} C(S)$.

Let us now work in the islander model in which there are n atomic propositions $A_1, \ldots, A_n$ and n knowledge agents $K_1, \ldots, K_n$. To model the statement that "it is commonly known that each islander knows each other islander's eye colour", one can use the sets of sentences

$$C(A_i \implies K_j(A_i)) \tag{1.2}$$

and

$$C(\neg A_i \implies K_j(\neg A_i)) \tag{1.3}$$

for all distinct $i, j \in \{1, \ldots, n\}$.

For any $0 \leq l \leq n$, let $B_{\geq l}$ denote the sentence that there are at least l blue-eyed islanders; this can be encoded as a suitable finite combination of the $A_1, \ldots, A_n$. For instance, $B_{\geq 0}$ can be expressed by any tautology, $B_{\geq 1}$ can be expressed by $A_1 \vee \ldots \vee A_n$, $B_{\geq n}$ can be expressed by $A_1 \wedge \ldots \wedge A_n$, and intermediate $B_{\geq l}$ can be expressed by more complicated formulae. Let B_l denote the statement that there are *exactly* k blue-eyed islanders; for instance, if $n = 3$, then B_1 can be expressed as

$$(A_1 \wedge \neg A_2 \wedge \neg A_3) \vee (\neg A_1 \wedge A_2 \wedge \neg A_3) \vee (\neg A_1 \wedge \neg A_2 \wedge A_3).$$

The following theorem asserts, roughly speaking, that if there are m blue-eyed islanders, and it is commonly known that there are at least l blue-eyed islanders, then all blue-eyed islanders can deduce their own eye colour if $m \leq l$, but not otherwise.

Theorem 1.5.6. *Let $\mathcal{T}$ be the set of sentences consisting of the union of* (1.2) *and* (1.3) *for all distinct $i, j \in \{1, \ldots, n\}$. Let $0 \leq m, l \leq n$. Let S denote the sentence*

$$S = \bigwedge_{i=1}^{n} (A_i \implies K_i(A_i))$$

(informally, S asserts that all blue-eyed islanders know their own eye colour).

(1) *If $m \leq l$, then*

$$\mathcal{T}, B_m, C(B_{\geq l}) \vdash_{L_\infty} S.$$

(2) *If $m > l$, then*

$$\mathcal{T}, B_m, C(B_{\geq l}) \nvdash_{L_\infty} S.$$

Proof. The first part of the theorem can be established informally as follows: if B_m holds, then each blue-eyed islander sees $m-1$ other blue-eyed islanders, but also knows that there are at least l blue-eyed islanders. If $m \leq l$, this forces each blue-eyed islander to conclude that his or her own eyes are blue (and in fact if $m < l$, the blue-eyed islander's knowledge is now inconsistent, but the conclusion is still valid thanks to *ex falso quodlibet*). It is a routine matter to formalise this argument using the axioms (1.2), (1.3) and the epistemic inference rule; we leave the details as an exercise.

To prove the second part, it suffices (by soundness) to construct a L_∞-model M_∞ which satisfies $\mathcal{T}$, B_m, and $C(B_{\geq l})$ but not S. By definition of an L_∞-model, it thus suffices to construct, for all sufficiently large natural numbers k, an L_∞-model M_k which satisfies $\mathcal{T} \cap L_k$, B_m, and $C(B_{\geq l}) \cap L_k$, but not S, and which are consistent with each other in the sense that each M_k is the restriction of M_{k+1} to L_k.

We can do this by a modification of the construction in Example 1.5.2. For any L_0-model M_0, we can recursively define an L_k-model $M_{k,\geq l}(M_0)$ for any $k \geq 0$ by setting $M_{0,\geq l}(M_0) := M_0$, and then for each $k \geq 1$, setting $M_{k,\geq l}(M_0)$ to be the L_k-model with L_0-model M_0, and with possible worlds $\mathcal{M}^{(i)}_{k-1}$ given by

$$\mathcal{M}^{(i)}_{k-1} := \{M_{k-1,\geq l}(M_{0,i}) : M_{0,i} \in \{M^+_{0,i}, M^-_{0,i}\}; M_{0,i} \models_{L_0} B_{\geq l}\};$$

this is the same construction as in Example 1.5.2, except that at all levels of the recursive construction, we restrict attention to worlds that obey $B_{\geq l}$. A routine induction shows that the $M_{k,\geq l}(M_0)$ determine a limit $M_{\infty,\geq l}(M_0)$, which is an L_∞ model that obeys $\mathcal{T}$ and $C(B_{\geq l})$. If $M_0 \models_{L_0} B_m$, then clearly $M_{\infty,\geq l}(M_0) \models_{L_\infty} B_m$ as well. But if $m > l$, then we see that $M_{\infty,\geq l}(M_0) \not\models_{L_\infty} S$, because for any index i with $M_0 \models_{L_0} A_i$, we see that if $k-1$, then $\mathcal{M}^{(i)}_{k-1}(M_0)$ contains worlds in which A_i is false, and so $M_{k,\geq l}(M_0) \not\models_{L_k} K_i(A_i)$ for any $k \geq 1$. □

1.5.2. Temporal epistemic logic. The epistemic logic discussed above is sufficiently powerful to model the knowledge environment of the islanders in the blue-eyed islander puzzle at a *single instant in time*, but in order to fully model the islander puzzle, we now must now incorporate the role of time. To avoid confusion, I feel that this is best accomplished by adopting a "spacetime" perspective, in which time is treated as another coordinate

rather than having any particularly privileged role in the theory, and the model incorporates all time-slices of the system at once. In particular, if we allow the time parameter t to vary along some set T of times, then each actor I_i in the model should now generate not just a single knowledge agent K_i, but instead a *family* $(K_{i,t})_{t\in T}$ of knowledge agents, one for each time $t \in T$. Informally, $K_{i,t}(S)$ should then model the assertion that "I_i knows S at time t". This of course leads to many more knowledge agents than before; if for instance one considers an islander puzzle with n islanders over M distinct points in time, this would lead to nM distinct knowledge agents $K_{i,t}$. And if the set of times T is countably or uncountably infinite, then the number of knowledge agents would similarly be countably or uncountably infinite. Nevertheless, there is no difficulty extending the previous epistemic logics L_k and L_∞ to cover this situation. In particular we still have a complete and sound logical framework to work in.

Note that if we do so, we allow for the ability to nest knowledge operators at different times in the past or future. For instance, if we have three times $t_1 < t_2 < t_3$, one could form a sentence such as

$$K_{1,t_2}(K_{2,t_1}(S)),$$

which informally asserts that at time t_2, I_1 knows that I_2 already knew S to be true by time t_1, or

$$K_{1,t_2}(K_{2,t_3}(S)),$$

which informally asserts that at time t_2, I_1 knows that I_2 will know S to be true by time t_3. The ability to know certain statements about the future is not too relevant for the blue-eyed islander puzzle, but is a crucial point in the unexpected hanging paradox.

Of course, with so many knowledge agents present, the models become more complicated; a model M_k of L_k now must contain inside it clouds $\mathcal{M}_{k-1}^{(i,t)}$ of possible worlds for each actor I_i and each time $t \in T$.

One reasonable axiom to add to a temporal epistemological system is the ability of agents to remember what they know. More precisely, we can impose the "memory axiom"

$$C(K_{i,t}(S) \implies K_{i,t'}(S)) \tag{1.4}$$

for any $S \in L_\infty$, any $i = 1, \ldots, m$, and any $t < t'$. (This axiom is important for the blue-eyed islander puzzle, though it turns out not to be relevant for the unexpected hanging paradox.)

We can also define a notion of common knowledge at a single time $t \in T$: given a sentence $S \in L_\infty$, we let $C_t(S)$ denote the set of sentences of the form

$$K_{i_1,t}(K_{i_2,t}(\ldots(K_{i_k,t}(S))\ldots))$$

where $k \geq 0$ and $i_1, \ldots, i_k \in \{1, \ldots, n\}$. This is a subset of $C(S)$, which is the set of all sentences of the form

$$K_{i_1,t_1}(K_{i_2,t_2}(\ldots(K_{i_k,t_k}(S))\ldots))$$

where $t_1, \ldots, t_k \in T$ can vary arbitrarily in T.

1.5.3. The blue-eyed islander puzzle. Now we can model the blue-eyed islander puzzle. To simplify things a bit, we will work with a discrete set of times $T = \mathbf{Z}$ indexed by the integers, with 0 being the day in which the foreigner speaks, and any other time t being the time t days after (or before, if t is negative) the foreigner speaks. (One can also work with a continuous time with only minor changes.) Note the presence of negative time; this is to help resolve the question (which often comes up in discussion of this puzzle) as to whether the islanders would already have committed suicide even before the foreigner speaks.

Also, the way the problem is set up, we have the somewhat notationally annoying difficulty that once an islander commits suicide, it becomes meaningless to ask whether that islander continues to know anything or not. To resolve this problem, we will take the liberty of modifying the problem by replacing "suicide" with a non-lethal public ritual. (This means (thanks to (1.4)) that once an islander learns his or her own eye colour, he or she will be condemned to repeating this ritual "suicide" every day from that point.) It is possible to create a logic which tracks when different agents are alive or dead and to thus model the concept of suicide, but this is something of a distraction from the key point of the puzzle, so we will simply redefine away this issue.

For similar reasons, we will not concern ourselves with eye colours other than blue, and only consider suicides stemming from blue eyes, rather than from any non-blue colour. (It is intuitively obvious, and can eventually be proven, that the foreigner's statement about the existence of blue-eyed islanders is insufficient information to allow any islander to distinguish between, say, green eyes and brown eyes, and so this statement cannot trigger the suicide of any non-blue-eyed person.)

As in previous sections, our logic will have the atomic propositions $A_1, \ldots, A_n$, with each A_i expressing the statement that I_i has blue eyes, as well as knowledge agents $K_{i,t}$ for each $i = 1, \ldots, n$ and $t \in \mathbf{Z}$. However, we will also need further atomic propositions $S_{i,t}$ for $i = 1, \ldots, n$ and $t \in \mathbf{Z}$, which denote the proposition that I_i commits suicide (or a ritual equivalent) at time t. Thus we now have a countably infinite number of atomic propositions and a countably infinite number of knowledge agents, but there is little difficulty extending the logics L_k and L_∞ to cover this setting.

We can now set up the various axioms for the puzzle. The "highly logical" axiom has already been subsumed in the epistemological inference rule. We also impose the memory axiom (1.4). Now we formalise the other assumptions of the puzzle:

- (All islanders see each other's eye colour) If $i, j \in \{1, \dots, n\}$ are distinct and $t \in \mathbf{Z}$, then

$$C(A_i \implies K_{j,t}(A_i)) \tag{1.5}$$

and

$$C(\neg A_i \implies K_{j,t}(\neg A_i)). \tag{1.6}$$

- (Anyone who learns their own eye colour is blue, must commit suicide the next day) If $i \in \{1, \dots, n\}$ and $t \in \mathbf{Z}$, then

$$C(K_{i,t}(A_i) \implies S_{i,t+1}). \tag{1.7}$$

- (Suicides are public) For any $i \in \{1, \dots, n\}$, $t \in \mathbf{Z}$, and $S'_{i,t} \in C_t(S_{i,t})$, we have

$$C(S_{i,t} \implies S'_{i,t}). \tag{1.8}$$

Similarly, if $S''_{i,t} \in C_t(\neg S_{i,t})$, then

$$C(\neg S_{i,t} \implies S''_{i,t}). \tag{1.9}$$

- (Foreigner announces in public on day 0 that there is at least one blue-eyed islander) We have

$$C_0(B_{\geq 1}). \tag{1.10}$$

Let $\mathcal{T}$ denote the union of all the axioms (1.4), (1.5), (1.6), (1.7), (1.8), (1.10). The "solution" to the islander puzzle can then be summarised as follows:

Theorem 1.5.7. *Let $1 \leq m \leq n$.*

(1) *(At least one blue-eyed islander commits suicide by day m)*

$$\mathcal{T}, B_m \vdash_{L_\infty} \bigvee_{i=1}^{n} \bigvee_{t=1}^{m} (A_i \wedge S_{i,t}).$$

(2) *(Nobody needs to commit suicide before day m) For any $t < m$ and $1 \leq i \leq m$,*

$$\mathcal{T}, B_m \nvdash_{L_\infty} S_{i,t}.$$

Note that the first conclusion is weaker than the conventional solution to the puzzle, which asserts in fact that all m blue-eyed islanders will commit suicide on day m. While this indeed the "default" outcome of the hypotheses $\mathcal{T}, B_m$, it turns out that this is not the only possible outcome; for instance, if one blue-eyed person happens to commit suicide on day 0 or day 1 (perhaps

for an unrelated reason than learning his or her own eye colour), then it turns out that this "cancels" the effect of the foreigner's announcement, and prevents further suicides. (So, if one were truly nitpicky, the conventional solution is not always correct, though one could also find similar loopholes to void the solution to most other logical puzzles, if one tried hard enough.)

In fact there is a strengthening of the first conclusion: given the hypotheses $\mathcal{T}, B_m$, there must exist a time $1 \leq t \leq m$ and t distinct islanders $I_{i_1}, \ldots, I_{i_t}$ such that $A_{i_j} \wedge S_{i_j,t}$ holds for all $j = 1, \ldots, t$.

Note that the second conclusion does not prohibit the existence of some models of $\mathcal{T}, B_m$ in which suicides occur before day m (consider for instance a situation in which a second foreigner made a similar announcement a few days before the first one, causing the chain of events to start at an earlier point and leading to earlier suicides).

Proof. (Sketch) To illustrate the first part of the theorem, we focus on the simple case $m = n = 2$; the general case is similar but requires more notation (and an inductive argument). It suffices to establish that

$$\mathcal{T}, B_2, \neg S_{1,1}, \neg S_{2,1} \vdash_{L_\infty} S_{1,2} \wedge S_{2,2}$$

(i.e. if nobody suicides by day 1, then both islanders will suicide on day 2.)

Assume $\mathcal{T}, B_2, \neg S_{1,1}, \neg S_{2,1}$. From (1.10) we have

$$K_{1,0}(K_{2,0}(A_1 \vee A_2))$$

and hence by (1.4)

$$K_{1,1}(K_{2,0}(A_1 \vee A_2)).$$

By (1.6) we also have

$$K_{1,1}(\neg A_1 \implies K_{2,0}(\neg A_1))$$

whereas from the epistemic inference axioms we have

$$K_{1,1}((K_{2,0}(A_1 \vee A_2) \wedge K_{2,0}(\neg A_1)) \implies K_{2,0}(A_2)).$$

From the epistemic inference axioms again, we conclude that

$$K_{1,1}(\neg A_1 \implies K_{2,0}(A_2))$$

and hence by (1.7) (and epistemic inference)

$$K_{1,1}(\neg A_1 \implies S_{2,1}).$$

On the other hand, from $\neg S_{2,1}$ and (1.9) we have

$$K_{1,1}(\neg S_{2,1})$$

and hence by epistemic inference

$$K_{1,1}(A_1)$$

and thus by (1.7)

$$S_{1,2}.$$

A similar argument gives $S_{2,2}$, and the claim follows.

To prove the second part, one has to construct, for each k, an L_k-model in which $\mathcal{T}, B_m$ is true and $S_{i,t}$ is false for any $1 \leq i \leq n$ and $t < m$. This is remarkably difficult, in large part due to the ability of nested knowledge operators to jump backwards and forwards in time. In particular, one can jump backwards to before Day 0, and so one must first model worlds in which there is no foreigner announcement. We do this as follows. Given an L_0-model M_0, we recursively define a L_k-model $M_k(M_0)$ for $k = 0, 1, 2, \ldots$ as follows. Firstly, $M_0(M_0) := M_0$. Next, if $k \geq 1$ and $M_{k-1}()$ has already been defined, we define $M_k(M_0)$ to be the L_k-model with L_0-model M_0, and for any $i = 1, \ldots, n$ and $t \in \mathbf{Z}$, setting $\mathcal{M}_{k-1}^{(i,t)}(M_0)$ to be the set of all L_{k-1}-models of the form $M_{k-1}(M_0')$, where M_0' is an L_0-model obeying the following properties:

- (I_i sees other islanders' eyes) If $j \in \{1, \ldots, n\}$ and $j \neq i$, then $M_0' \models_{L_0} A_i$ iff $M_0 \models_{L_0} A_i$.
- (I_i remembers suicides) If $j \in \{1, \ldots, n\}$ and $t' \leq t$, then $M_0' \models_{L_0} S_{j,t'}$ iff $M_0 \models_{L_0} S_{j,t'}$.

Now we model worlds in which there is a foreigner announcement. Define an *admissible L_0 model* to be an L_0-model M_0 such that there exist $1 \leq t \leq m$ for which the following hold:

- $M_0 \models_{L_0} B_m$ (i.e. there are exactly m blue-eyed islanders in the world M_0).
- There exists distinct $i_1, \ldots, i_t \in \{1, \ldots, n\}$ such that $M_0 \models_{L_0} A_{i_j}$ and $M_0 \models_{L_0} S_{i_j,t}$ for all $j = 1, \ldots, t$.
- For any $i \in \{1, \ldots, n\}$ and $t' \in \mathbf{Z}$, $M_0 \models_{L_0} S_{i,t'}$ implies $M_0 \models_{L_0} S_{i,t'+1}$.

We call m the *blue-eyed count* of M_0.

(Such models, incidentally, can already be used to show that no suicides necessarily occur in the absence of the foreigner's announcement, because the limit $M_\infty(M_0)$ of such models always obey all the axioms of $\mathcal{T}$ except for (1.10).)

Given an admissible L_0-model M_0 of some blue-eyed count m, we recursively define an L_k model $\tilde{M}_k(M_0)$ for $k = 0, 1, 2, \ldots$ by setting $\tilde{M}_0(M_0) := M_0$, then if $k \geq 1$ and $\tilde{M}_{k-1}()$ has already been defined, we define $\tilde{M}_k(M_0)$ to be the L_k-model with L_0-model M_0, and with $\mathcal{M}_{k-1}^{(i,t)}(M_0)$ for $i = 1, \ldots, n$ and $t \in \mathbf{Z}$ defined by the following rules:

Case 1. If $t < 0$, then we set $\mathcal{M}_{k-1}^{(i,t)}(M_0)$ to be the set of all L_{k-1}-models of the form $M_{k-1}(M_0')$, where M_0' obeys the two properties "I_i sees other islanders' eyes" and "I_i remembers suicides" from the preceding construction. (M_0' does not need to be admissible in this case.)

Case 2. If $t = m - 1$, $M_0 \models A_i$, and there does not exist $1 \leq t' \leq t$ distinct $i_1, \ldots, i_{t'} \in \{1, \ldots, n\}$ such that $M_0 \models_{L_0} A_{i_j} \wedge S_{i_j,t'}$ for all $j = 1, \ldots, t'$, then we set $\mathcal{M}_{k-1}^{(i,t)}(M_0)$ L_{k-1}-models of the form $\tilde{M}_{k-1}(M_0')$, where M_0' is admisssible, obeys the two properties "I_i sees other islanders' eyes" and "I_i remembers suicides" from the preceding construction, and also obeys the additional property $M_0' \models A_i$. (Informally, this is the case in which I_i "must learn" A_i.)

Case 3. In all other cases, we set $\mathcal{M}_{k-1}^{(i,t)}(M_0)$ to be the set of all L_{k-1}-models of the form $\tilde{M}_{k-1}(M_0')$, where M_0' is admissible and obeys the two properties "I_i sees other islanders' eyes" and "I_i remembers suicides" from the preceding construction.

We let $\tilde{M}_\infty(M_0)$ be the limit of the $\tilde{M}_k(M_0)$ (which can easily be verified to exist by induction). A quite tedious verification reveals that for any admissible L_0-model M_0 of blue-eyed count m, that $\tilde{M}_\infty(M_0)$ obeys both $\mathcal{T}$ and B_m, but one can choose M_0 to not admit any suicides before time m, which will give the second claim of the theorem. □

Remark 1.5.8. Under the assumptions used in our analysis, we have shown that it is inevitable that the foreigner's comment will cause at least one death. However, it is possible to avert all deaths by breaking one or more of the assumptions used. For instance, if it is possible to sow enough doubt in the islanders' minds about the logical and devout nature of the other islanders, then one can cause a breakdown of the epistemic inference rule or of (1.7), and this can prevent the chain of deductions from reaching its otherwise deadly conclusion.

Remark 1.5.9. The same argument actually shows that L_∞ can be replaced by L_m for the first part of Theorem 1.5.7 (after restricting the definition of common knowledge to those sentences that are actually in L_m, of course). On the other hand, using L_k for $k < m$, one can show that this logic is insufficient to deduce any suicides if there are m blue-eyed islanders, by using the model $M_k(M_0)$ defined above; we omit the details.

1.5.4. The unexpected hanging paradox. We now turn to the *unexpected hanging paradox*, and try to model it using (temporal) epistemic logic. Here is a common formulation of the paradox (taken from the Wikipedia entry on this problem):

Problem 1.5.10. A judge tells a condemned prisoner that he will be hanged at noon on one weekday in the following week but that the execution will be a surprise to the prisoner. He will not know the day of the hanging until the executioner knocks on his cell door at noon that day.

Having reflected on his sentence, the prisoner draws the conclusion that he will escape from the hanging. His reasoning is in several parts. He begins by concluding that the "surprise hanging" can't be on Friday, as if he hasn't been hanged by Thursday, there is only one day left - and so it won't be a surprise if he's hanged on Friday. Since the judge's sentence stipulated that the hanging would be a surprise to him, he concludes it cannot occur on Friday.

He then reasons that the surprise hanging cannot be on Thursday either, because Friday has already been eliminated and if he hasn't been hanged by Wednesday night, the hanging must occur on Thursday, making a Thursday hanging not a surprise either. By similar reasoning he concludes that the hanging can also not occur on Wednesday, Tuesday or Monday. Joyfully he retires to his cell confident that the hanging will not occur at all. The next week, the executioner knocks on the prisoner's door at noon on Wednesday which, despite all the above, was an utter surprise to him. Everything the judge said came true.

It turns out that there are several, not quite equivalent, ways to model this "paradox" epistemologically, with the differences hinging on how one interprets what "unexpected" or "surprise" means. In particular, if S is a sentence and K is a knowledge agent, how would one model the sentence "K does not expect S" or "K is surprised by S"?

One possibility is to model this sentence as

$$\neg K(S), \tag{1.11}$$

i.e. as the assertion that K does not know S to be true. However, this leads to the following situation: if K has inconsistent knowledge (in particular, one has $K(\bot)$, where $\bot$ represents falsity (the negation of a tautology)), then by *ex falso quodlibet*, $K(S)$ would be true for every S, and hence K would expect everything and be surprised by nothing. An alternative interpretation, then, is to adopt the convention that an agent with inconsistent knowledge is so confused as to not be capable of expecting anything (and thus be surprised by everything). In this case, "K does not expect S" should instead be modeled as

$$(\neg K(S)) \vee K(\bot), \tag{1.12}$$

i.e. that K either does not know S to be true, or is inconsistent.

Both interpretations (1.11), (1.12) should be compared with the sentence

$$K(\neg S), \tag{1.13}$$

i.e. that K knows that S is false. If K is consistent, (1.13) implies (1.11), but if K is inconsistent then (1.13) is true and (1.11) is false. In either case, though, we see that (1.13) implies (1.12).

Let now analyse the unexpected hanging paradox using the former interpretation (1.11) of surprise. We begin with the simplest (and somewhat degenerate) situation, in which there is only one time (say Monday at noon) in which the hanging is to take place. In this case, there is just one knowledge agent K (the knowledge of the prisoner after the judge speaks, but before the executation date of Monday at noon). We introduce an atomic sentence E, representing the assertion that the prisoner will be hanged on Monday at noon. In this case (and using the former interpretation (1.11) of surprise), the judge's remarks can be modeled by the sentence

$$S := E \wedge (E \implies \neg K(E)).$$

The "paradox" in this case stems from the following curious fact:

Theorem 1.5.11. (1) *There exist L_∞-models in which S is true.*

(2) *There exist L_∞-models in which $K(S)$ is true.*

(3) *However, there does not exist any L_∞-model in which both S and $K(S)$ is true.*

Thus, the judge's statement can be true, but if so, it is not possible for the prisoner to know this! (In this regard, the sentence S is analogous to a Godel sentences, which can be true in models of a formal system, but not provable in that system.) More informally: knowing a surprise, ruins that surprise.

Proof. The third statement is easy enough to establish: if S is true in some model, then clearly $\neg K(E)$ is true in that model; but if $K(S)$ is true in the same model, then (by epistemic inference) $K(E)$ will be true as well, which is a contradiction.

The first statement is also fairly easy to establish. We have two L_0-models; a model M_0^+ in which E is true, and a model M_0^- in which E is false. We can recursively define the L_k-model $M_k(M_0)$ for any $k \geq 0$ and any $M_0 \in \{M_0^+, M_0^-\}$ by setting $M_0(M_0) := M_0$, and for $k \geq 1$, setting $M_k(M_0)$ to be the L_k-model with L_0-model M_0, and with $\mathcal{M}_{k-1} := \{M_{k-1}(M_0^+), M_{k-1}(M_0^-)\}$. One then easily verifies that the $M_k(M_0)$ have a limit $M_\infty(M_0)$, and that $M_\infty(M_0^+)$ models S (but not $K(S)$, of course).

A trivial way to establish the second statement is to make a model in which K is inconsistent (thus $\mathcal{M}_{k-1}$ is empty). One can also take $\mathcal{M}_{k-1}$ to

be $M_{k-1}(M_0^+)$, and this will also work. (Of course, in such models, S must be false.) □

Another peculiarity of the sentence S is that

$$K(S), K(K(S)) \models_{L_\infty} K(\bot)$$

as can be easily verified (by modifying the proof of the second statement of the above theorem). Thus, the sentence S has the property that if the prisoner believes S, and also knows that he or she believes S, then the prisoner's beliefs automatically become inconsistent - despite the fact that S is not actually a self-contradictory statement (unless also combined with $K(S)$).

Now we move to the case when the execution could take place at two possible times, say Monday at noon and Tuesday at noon. We then have two atomic statements: E_1, the assertion that the execution takes place on Monday at noon, and E_2, the assertion that the execution takes place on Tuesday at noon. There are two knowledge agents; K_1, the state of knowledge just before Monday at noon, and K_2, the state of knowledge just before Tuesday at noon. (There is again the annoying notational issue that if E_1 occurs, then presumably the prisoner will have no sensible state of knowledge by Tuesday, and so K_2 might not be well defined in that case; to avoid this irrelevant technicality, we replace the execution by some non-lethal punishment (or use an alternate formulation of the puzzle, for instance by replacing an unexpected hanging with a surprise exam.)

We will need one axiom beyond the basic axioms of epistemic logic, namely

$$C(\neg E_1 \implies K_2(\neg E_1)). \tag{1.14}$$

Thus, it is common knowledge that if the execution does not happen on Monday, then by Tuesday, the prisoner will be aware of this fact. This axiom should of course be completely non-controversial.

The judge's sentence, in this case, is given by

$$S := (E_1 \vee E_2) \wedge (E_1 \implies \neg K_1(E_1)) \wedge (E_2 \implies \neg K_2(E_2)).$$

Analogously to Theorem 1.5.11, we can find L_∞ models obeying (1.14) in which S is true, but one cannot find models obeying (1.14) in which S, $K_1(S)$, $K_2(S)$, and $K_1(K_2(S))$ are all true, as one can soon see that this leads to a contradiction. Indeed, from S one has

$$\neg E_1 \implies \neg K_2(E_2)$$

while from (1.14) one has

$$\neg E_1 \implies K_2(\neg E_1)$$

and from $K_2(S)$ one has

$$K_2(E_1 \wedge E_2)$$

wihch shows that $\neg E_1$ leads to a contradiction, which implies E_1 and hence $\neg K_1(E_1)$ by S. On the other hand, from $K_1(S)$ one has

$$K_1(\neg E_1 \implies \neg K_2(E_2))$$

while from (1.14) one has

$$K_1(\neg E_1 \implies K_2(\neg E_1))$$

and from $K_1(K_2(S))$ one has

$$K_1(K_2(E_1 \wedge E_2))$$

which shows that $K_1(\neg E_1 \implies \bot)$, and thus $K_1(E_1)$, a contradiction. So, as before, S is a secret which can only be true as long as it is not too widely known.

A slight variant of the above argument shows that if $K_1(S)$, $K_2(S)$, and $K_1(K_2(S))$ hold, then $K_1(\neg E_1)$ and $\neg E_1 \implies K_2(\neg E_2)$ hold - or informally, the prisoner can deduce using knowledge of S (and knowledge of knowledge of S) that there will be no execution on either date. This may appear at first glance to be consistent with S (which asserts that the prisoner will be surprised when the execution does happen), but this is a confusion between (1.11) and (1.13). Indeed, one can show under the assumptions $K_1(S), K_2(S), K_1(K_2(S))$ that K_1 is inconsistent, and (if $\neg E_1$ holds) then K_2 is also inconsistent, and so $K_1(\neg E_1)$ and $\neg E_1 \implies K_2(\neg E_2)$ do not, in fact, imply S.

Now suppose that we interpret surprise using (1.12) instead of (1.11). Let us begin first with the one-day setting. Now the judge's sentence becomes

$$S = E \wedge (E \implies (\neg K(E) \vee K(\bot))).$$

In this case it is possible for S and $K(S)$ to be true, and in fact for S to be common knowledge, basically by making K inconsistent. (A little more precisely: we use the L_k-model M_k where $M_0 = M_0^+$ and $\mathcal{M}_{k-1} = \emptyset$. Informally: the judge has kept the execution a surprise by driving the prisoner insane with contradictions.

The situation is more interesting in the two-day setting (as first pointed out by Kritchman and Raz [**KrRa2010**]), where S is now

$$S := (E_1 \vee E_2) \wedge (E_1 \implies (\neg K_1(E_1) \vee K_1(\bot)))$$
$$\wedge (E_2 \implies (\neg K_2(E_2) \vee K_2(\bot))).$$

Here it is possible for S to in fact be common knowledge in some L_∞-model, but in order for this to happen, at least one of the following three statements must be true in this model:

- $K_1(\bot)$.
- $K_2(\bot)$.
- $\neg K_1(\neg K_2(\bot))$.

(We leave this as an exercise for the interested reader.) In other words, in order for the judge's sentence to be common knowledge, either the prisoner's knowledge on Monday or Tuesday needs to be inconsistent, or else the prisoner's knowledge is consistent, but the prisoner is unable (on Monday) to determine that his or her own knowledge (on Tuesday) is consistent. Notice that the third conclusion here $\neg K_1(\neg K_2(\bot))$ is very reminiscent of *Gödel's second incompleteness theorem*, and indeed in [**KrRa2010**], the surprise examination argument is modified to give a rigorous proof of that theorem.

Remark 1.5.12. Here is an explicit example of a L_∞-world in which S is common knowledge, and K_1 and K_2 are both consistent (but K_1 does not know that K_2 is consistent). We first define L_k-models M_k for each $k = 0, 1, \ldots$ recursively by setting M_0 to be the world in which $M_0 \models_{L_0} E_2$ and $M_0 \models_{L_0} \neg E_1$, and then define M_k for $k \geq 1$ to be the L_k-model with L_0-model M_0, with $\mathcal{M}^{(1)}_{k-1} := \{M_{k-1}\}$, and $\mathcal{M}^{(2)}_{k-1} := \emptyset$. (Informally: the execution is on Tuesday, and the prisoner knows this on Monday, but has become insane by Tuesday.) We then define the models $\tilde{M}_k$ for $k = 0, 1, \ldots$ recursively by setting $\tilde{M}_0$ to be the world in which $\tilde{M}_0 \models_{L_0} E_1$ and $\tilde{M}_0 \models_{L_0} \neg E_2$, then define $\tilde{M}_k$ for $k \geq 1$ to be the L_k-model with L_0-model $\tilde{M}_0$, $\mathcal{M}^{(1)}_{k-1} := \{M_{k-1}, \tilde{M}_{k-1}\}$, and $\mathcal{M}^{(2)}_{k-1} := \{\tilde{M}_{k-1}\}$. (Informally: the execution is on Monday, but the prisoner only finds this out after the fact.) The limit $\tilde{M}_\infty$ of the $\tilde{M}_k$ then has S as common knowledge, with $K_1(\bot)$ and $K_2(\bot)$ both false, but $K_1(\neg K_2(\bot))$ is also false.

Chapter 2

Group theory

2.1. Symmetry spending

Many problems in mathematics have the general form "For any object x in the class X, show that the property $P(x)$ is true". For instance, one might need to prove an identity or inequality for all choices of parameters x (which may be numbers, functions, sets, or other objects) in some parameter space X.

In many cases, such problems enjoy invariance or closure properties with respect to some natural symmetries, actions, or operations. For instance, there might be an operation T that preserves X (so that if x is in X, then Tx is in X) and preserves P (so that if $P(x)$ is true, then $P(Tx)$ is true). Then, in order to verify the problem for Tx, it suffices to verify the problem for x.

Similarly, if X is closed under (say) addition, and P is also closed under addition (thus if $P(x)$ and $P(y)$ is true, then $P(x+y)$ is true), then to verify the problem for $x+y$, it suffices to verify the problem for x and y separately.

Another common example of a closure property: if X is closed under some sort of limit operation, and P is also closed under the same limit operation (thus if x_n converges to x and $P(x_n)$ is true for all n, then $P(x)$ is true), then to verify the problem for x, then it suffices to verify the problem for the x_n.

One can view these sorts of invariances and closure properties as problem-solving *assets*; in particular, one can *spend* these assets to reduce the class X of objects x that one needs to solve the problem for. By doing so, one has to give up the invariance or closure property that one spent; but if one spends these assets wisely, this is often a favorable tradeoff. (And one can often

buy back these assets if needed by expanding the class of objects again (and defining the property P in a sufficiently abstract and invariant fashion).)

For instance, if one needs to verify $P(x)$ for all x in a normed vector space X, and the property $P(x)$ is homogeneous (so that, for any scalar c, $P(x)$ implies $P(cx)$), then we can spend this homogeneity invariance to normalise x to have norm 1, thus effectively replacing X with the unit sphere of X. Of course, this new space is no longer closed under homogeneity; we have spent that invariance property. Conversely, to prove a property $P(x)$ for all x on the unit sphere, it is equivalent to prove $P(x)$ for all x in X, provided that one extends the definition of P(x) to X in a homogeneous fashion.

As a rule of thumb, each independent symmetry of the problem that one has can be used to achieve one normalisation. Thus, for instance, if one has a three-dimensional group of symmetries, one can expect to normalise three quantities of interest to equal a nice value (typically one normalises to 0 for additive symmetries, or 1 for multiplicative symmetries).

In a similar spirit, if the problem one is trying to solve is closed with respect to an operation such as addition, then one can restrict attention to all x in a suitable generating set of X, such as a basis. Many "divide and conquer" strategies are based on this type of observation.

Or: if the problem one is trying to solve is closed with respect to limits, then one can restrict attention to all x in a dense subclass of X. This is a particularly useful trick in real analysis (using limiting arguments to replace reals with rationals, sigma-compact sets with compact sets, rough functions with nice functions, etc.). If one uses ultralimits instead of limits, this type of observation leads to various useful correspondence principles between finitary instances of the problem and infinitary ones (with the former serving as a kind of "dense subclass" of the latter); see e.g. [**Ta2012**, §1.7].

Sometimes, one can exploit rather abstract or unusual symmetries. For instance, certain types of statements in algebraic geometry tend to be insensitive to the underlying field (particularly if the fields remain algebraically closed). This allows one to sometimes move from one field to another, for instance from an infinite field to a finite one or vice versa; see [**Ta2010b**, §1.2]. Another surprisingly useful symmetry is closure with respect to tensor powers; see [**Ta2008**, §1.9].

Gauge symmetry is a good example of a symmetry which is both spent (via gauge fixing) and bought (by reformulating the problem in a gauge-invariant fashion); see [**Ta2009b**, §1.4].

Symmetries also have many other uses beyond their ability to be spent in order to obtain normalisation. For instance, they can be used to analyse a claim or argument for compatibility with that symmetry; generally speaking, one should not be able to use a non-symmetric argument to prove a symmetric claim (unless there is an explicit step where one spends the symmetry in a strategic fashion). The useful tool of dimensional analysis is perhaps the most familiar example of this sort of meta-analysis.

Thanks to Noether's theorem and its variants, we also know that there often is a duality relationship between (continuous) symmetries and conservation laws; for instance, the time-translation invariance of a (Hamiltonian or Lagrangian) system is tied to energy conservation, the spatial translation invariance is tied to momentum conservation, and so forth. The general principle of relativity (that the laws of physics are invariant with respect to arbitrary nonlinear coordinate changes) leads to a much stronger pointwise conservation law, namely the divergence-free nature of the stress-energy tensor, which is fundamentally important in the theory of wave equations (and particularly in general relativity).

As the above examples demonstrate, when solving a mathematical problem, it is good to be aware of what symmetries and closure properties the problem has, before one plunges in to a direct attack on the problem. In some cases, such symmetries and closure properties only become apparent if one abstracts and generalises the problem to a suitably "natural' framework; this is one of the major reasons why mathematicians use abstraction even to solve concrete problems. (To put it another way, abstraction can be used to purchase symmetries or closure properties by spending the implicit normalisations that are present in a concrete approach to the problem; see [**Ta2011d**, §1.6].)

2.2. Isogenies between classical Lie groups

For sake of concreteness we will work here over the complex numbers $\mathbf{C}$, although most of this discussion is valid for arbitrary algebraically closed fields (but some care needs to be taken in characteristic 2, as always, particularly when defining the orthogonal and symplectic groups). Then one has the following four infinite families of *classical Lie groups* for $n \geq 1$:

(1) (Type A_n) The *special linear group* $SL_{n+1}(\mathbf{C})$ of volume-preserving linear maps $T : \mathbf{C}^{n+1} \to \mathbf{C}^{n+1}$.

(2) (Type B_n) The *special orthogonal group* $SO_{2n+1}(\mathbf{C})$ of (orientation preserving) linear maps $T : \mathbf{C}^{2n+1} \to \mathbf{C}^{2n+1}$ preserving a non-degenerate symmetric form $\langle , \rangle : \mathbf{C}^{2n+1} \times \mathbf{C}^{2n+1} \to \mathbf{C}$, such as the

standard symmetric form

$$\langle (z_1, \ldots, z_{2n+1}), (w_1, \ldots, w_{2n+1}) \rangle := z_1 w_1 + \ldots + z_{2n+1} w_{2n+1}.$$

(this is the complexification of the more familiar *real special orthogonal group* $SO_{2n+1}(\mathbf{R})$).

(3) (Type C_n) The *symplectic group* $Sp_{2n}(\mathbf{C})$ of linear maps $T : \mathbf{C}^{2n} \to \mathbf{C}^{2n}$ preserving a non-degenerate antisymmetric form $\omega : \mathbf{C}^{2n} \times \mathbf{C}^{2n} \to \mathbf{C}$, such as the standard symplectic form

$$\omega((z_1, \ldots, z_{2n}), (w_1, \ldots, w_{2n})) := \sum_{j=1}^{n} z_j w_{n+j} - z_{n+j} w_j.$$

(4) (Type D_n) The special orthogonal group $SO_{2n}(\mathbf{C})$ of (orientation preserving) linear maps $\mathbf{C}^{2n} \to \mathbf{C}^{2n}$ preserving a non-degenerate symmetric form $\langle , \rangle : \mathbf{C}^{2n} \times \mathbf{C}^{2n} \to \mathbf{C}$ (such as the standard symmetric form).

In this section, we will abuse notation somewhat and identify A_n with $SL_{n+1}(\mathbf{C})$, B_n with $SO_{2n+1}(\mathbf{C})$, etc., although it is more accurate to say that $SL_{n+1}(\mathbf{C})$ is a Lie group of *type* A_n, etc., as there are other forms of the Lie algebras associated to A_n, B_n, C_n, D_n over various fields. Over a non-algebraically closed field, such as $\mathbf{R}$, the list of Lie groups associated with a given type can in fact get quite complicated, and will not be discussed here. One can also view the double covers $\mathrm{Spin}_{2n+1}(\mathbf{C})$ and $\mathrm{Spin}_{2n}(\mathbf{C})$ of $SO_{2n+1}(\mathbf{C})$, $SO_{2n}(\mathbf{C})$ (i.e. the *spin groups*) as being of type B_n, D_n respectively; however, I find the spin groups less intuitive to work with than the orthogonal groups and will therefore focus more on the orthogonal model.

The reason for this subscripting is that each of the classical groups A_n, B_n, C_n, D_n has *rank* n, i.e. the dimension of any maximal connected abelian subgroup of simultaneously diagonalisable elements (also known as a *Cartan subgroup*) is n. For instance:

(1) (Type A_n) In $SL_{n+1}(\mathbf{C})$, one Cartan subgroup is the diagonal matrices in $SL_{n+1}(\mathbf{C})$, which has dimension n.

(2) (Type B_n) In $SO_{2n+1}(\mathbf{C})$, all Cartan subgroups are isomorphic to $SO_2(\mathbf{C})^n \times SO_1(\mathbf{C})$, which has dimension n.

(3) (Type C_n) In $Sp_{2n}(\mathbf{C})$, all Cartan subgroups are isomorphic to $SO_2(\mathbf{C})^n \leq Sp_2(\mathbf{C})^n \leq Sp_{2n}(\mathbf{C})$, which has dimension n.

(4) (Type D_n) in $SO_{2n}(\mathbf{C})$, all Cartan subgroups are isomorphic to $SO_2(\mathbf{C})^n$, which has dimension n.

Remark 2.2.1. This same convention also underlies the notation for the *exceptional simple Lie groups* G_2, F_4, E_6, E_7, E_8, which we will not discuss further here.

With two exceptions, the classical Lie groups A_n, B_n, C_n, D_n are all *simple*, i.e. their Lie algebras are non-abelian and not expressible as the direct sum of smaller Lie algebras. The two exceptions are $D_1 = SO_2(\mathbf{C})$, which is abelian (isomorphic to $\mathbf{C}^\times$, in fact) and thus not considered simple, and $D_2 = SO_4(\mathbf{C})$, which turns out to "essentially" split as $A_1 \times A_1 = SL_2(\mathbf{C}) \times SL_2(\mathbf{C})$, in the sense that the former group is double covered by the latter (and in particular, there is an *isogeny* from the latter to the former, and the Lie algebras are isomorphic).

The *adjoint action* of a Cartan subgroup of a Lie group G on the Lie algebra $\mathfrak{g}$ splits that algebra into *weight spaces*; in the case of a simple Lie group, the associated weights are organised by a *Dynkin diagram*. The Dynkin diagrams for A_n, B_n, C_n, D_n are of course well known, and can be found in any text on Lie groups or algebraic groups.

For small n, some of these Dynkin diagrams are isomorphic; this is a classic instance of the tongue-in-cheek *strong law of small numbers* [**Gu1988**], though in this case "strong law of small diagrams" would be more appropriate. These accidental isomorphisms then give rise to the *exceptional isomorphisms* between Lie algebras (and thence to *exceptional isogenies* between Lie groups). Excluding those isomorphisms involving the exceptional Lie algebras E_n for $n = 3, 4, 5$, these isomorphisms are

(1) $A_1 = B_1 = C_1$;
(2) $B_2 = C_2$;
(3) $D_2 = A_1 \times A_1$;
(4) $D_3 = A_3$.

There is also a pair of exceptional isomorphisms from (the Spin_8 form of) D_4 to itself, a phenomenon known as *triality*.

These isomorphisms are most easily seen via algebraic and combinatorial tools, such as an inspection of the Dynkin diagrams. However, the isomorphisms listed above[1] can also be seen by more "geometric" means, using the basic representations of the classical Lie groups on their natural vector spaces ($\mathbf{C}^{n+1}, \mathbf{C}^{2n+1}, \mathbf{C}^{2n}, \mathbf{C}^{2n}$ for A_n, B_n, C_n, D_n respectively) and combinations thereof (such as *exterior powers*). These isomorphisms are quite standard (they can be found, for instance, in [**Pr2007**]), but I decided to present them here for sake of reference.

[1]However, I don't know of a simple way to interpret triality geometrically; the descriptions I have seen tend to involve some algebraic manipulation of the octonions or of a Clifford algebra, in a manner that tended to obscure the geometry somewhat.

2.2.1. $A_1 = C_1$**.** This is the simplest correspondence. $A_1 = SL_2(\mathbf{C})$ is the group of transformations $T : \mathbf{C}^2 \to \mathbf{C}^2$ that preserve the volume form; $C_1 = Sp_2(\mathbf{C})$ is the group of transformations $T : \mathbf{C}^2 \to \mathbf{C}^2$ that preserve the symplectic form. But in two dimensions, the volume form and the symplectic form are the same.

2.2.2. $A_1 = B_1$**.** The group $A_1 = SL_2(\mathbf{C})$ naturally acts on $\mathbf{C}^2$. But it also has an obvious three-dimensional action, namely the adjoint action $g : X \mapsto gXg^{-1}$ on the Lie algebra $\mathfrak{sl}_2(\mathbf{C})$ of 2×2 complex matrices of trace zero. This action preserves the Killing form

$$\langle X, Y \rangle_{\mathfrak{sl}_2(\mathbf{C})} := \mathrm{tr}(XY)$$

due to the cyclic nature of the trace. The Killing form is symmetric and non-degenerate (this reflects the simple nature of A_1), and so we see that each element of $SL_2(\mathbf{C})$ has been mapped to an element of

$$\mathrm{SO}(\mathfrak{sl}_2(\mathbf{C})) \equiv SO_3(\mathbf{C}) = B_1,$$

thus giving a homomorphism from A_1 to B_1. The group A_1 has dimension $2^2 - 1 = 3$, and B_1 has dimension $3(3-1)/2 = 3$, so A_1 and B_1 have the same dimension. The kernel of the map is easily seen to be the centre $\{+1, -1\}$ of A_1, and so this is a double cover[2] of B_1 by A_1 (thus interpreting $A_1 = SL_2(\mathbf{C})$ as the spin group $Spin_3(\mathbf{C})$).

A slightly different interpretation of this correspondence, using quaternions, will be discussed in Section 8.3.

2.2.3. $A_3 = D_3$**.** The group $A_3 = SL_4(\mathbf{C})$ naturally acts on $\mathbf{C}^4$. Like A_1, it has an adjoint action (on the 15-dimensional Lie algebra $\mathfrak{sl}_4(\mathbf{C})$), but this is not the action we will use for the $A_3 = D_3$ correspondence. Instead, we will look at the action on the $\binom{4}{2} = 6$-dimensional *exterior power* $\bigwedge^2 \mathbf{C}^4$ of $\mathbf{C}^4$, given by the usual formula

$$g(v \wedge w) := (gv) \wedge (gw).$$

Since $2+2 = 4$, the volume form on $\mathbf{C}^4$ induces a bilinear form $\langle , \rangle$ on $\bigwedge^2 \mathbf{C}^2$; since 2 is even, this form is symmetric rather than anti-symmetric, and it is also non-degenerate. An element of $SL_4(\mathbf{C})$ preserves the volume form and thus preserves the bilinear form, giving a map from $SL_4(\mathbf{C})$ to

$$\mathrm{SO}(\bigwedge^2 \mathbf{C}^4) \equiv SO_6(\mathbf{C}) = D_3.$$

This is a homomorphism from A_3 to D_3. The group A_3 has dimension $4^2 - 1 = 15$, and D_3 has dimension $6(6-1)/2 = 15$, so A_3 and D_3 have the same dimension. As before, the kernel is seen to be $\{+1, -1\}$, so this is a

[2]Note that the image of the map is open and B_1 is connected, so that one indeed has a covering map.

double cover of D_3 by A_3 (thus interpreting $A_3 = SL_4(\mathbf{C})$ as the spin group $\mathrm{Spin}_6(\mathbf{C})$).

2.2.4. $B_2 = C_2$**.** This is basically a restriction of the $A_3 = D_3$ correspondence. Namely, the group $C_2 = Sp_4(\mathbf{C})$ acts on $\mathbf{C}^4$ in a manner that preserves the symplectic form ω, and hence (on taking a wedge product) the volume form also. Thus C_2 is a subgroup of $SL_4(\mathbf{C}) = A_3$, and as discussed above, thus acts orthogonally on the six-dimensional space $\bigwedge^2 \mathbf{C}^4$. On the other hand, the symplectic form ω can itself be thought of as an element of $\bigwedge^2 \mathbf{C}^4$, and is clearly fixed by all of C_2; thus C_2 also stabilises the five-dimensional orthogonal complement $\omega^\perp$ of ω inside $\bigwedge^2 \mathbf{C}^4$. Note that ω is non-degenerate (here we crucially use the fact that the characteristic is not two!) and so $\omega^\perp$ is also non-degenerate. We have thus mapped C_2 to

$$\mathrm{SO}(\omega^\perp) \equiv SO_5(\mathbf{C}) = B_2.$$

This is a homomorphism from C_2 to B_2. The group C_2 has dimension $2(4+1) = 10$, while B_2 has dimension $5(5-1)/2 = 10$, so B_2 and C_2 have the same dimension. Once again, one can verify that the kernel is $\{+1, -1\}$, so this is a double cover of B_2 by C_2 (thus interpreting $C_2 = Sp_4(\mathbf{C})$ as the spin group $\mathrm{Spin}_5(\mathbf{C})$).

Remark 2.2.2. In characteristic two, the above map from C_2 to B_2 disappears, but there is a somewhat different identification between $B_n = SO_{2n+1}(k)$ and $C_n = Sp_{2n}(k)$ for any n in this case. Namely, in characteristic two, inside k^{2n+1} with a non-degenerate symmetric form $\langle , \rangle$, the set of null vectors (vectors x with $\langle x, x\rangle = 0$) forms a $2n$-dimensional hyperplane, and the restriction of the symmetric form to that hyperplane becomes a symplectic form (which, in characteristic two, is defined to be an anti-symmetric form ω with $\omega(x, x) = 0$ for all x). This provides the claimed identification between B_n and C_n.

2.2.5. $D_2 = A_1 \times A_1$**.** The group $A_1 \times A_1 = SL_2(\mathbf{C}) \times SL_2(\mathbf{C})$ acts on $\mathbf{C}^2 \times \mathbf{C}^2$ by direct sum:

$$(g, h)(v, w) := (gv, hw).$$

Each individual factor g, h preserves the symplectic form ω on $\mathbf{C}^2$, and so the pair (g, h) preserves the tensor product $\omega \otimes \omega$, which is the bilinear form on $\mathbf{C}^2 \times \mathbf{C}^2$ defined as

$$\omega \otimes \omega((v, w), (v', w')) := \omega(v, v')\omega(w, w').$$

As each factor ω is anti-symmetric and non-degenerate, the tensor product $\omega \otimes \omega$ is symmetric and non-degenerate. Thus we have mapped $A_1 \times A_1$ into

$$\mathrm{SO}(\mathbf{C}^2 \times \mathbf{C}^2) = SO_4(\mathbf{C}) = D_2.$$

The group $A_1 \times A_1$ has dimension $(2^2 - 1) + (2^2 - 1) = 6$, and D_2 has dimension $4(4-1)/2 = 6$, so $A_1 \times A_1$ and D_2 have the same dimension. As before, the kernel can be verified to be $\{(+1,+1),(-1,-1)\}$, and so this is a double cover of D_2 by $A_1 \times A_1$ (thus interpreting $A_1 \times A_1 = SL_2(\mathbf{C}) \times SL_2(\mathbf{C})$ as the spin group $\mathrm{Spin}_4(\mathbf{C})$).

Remark 2.2.3. All of these exceptional isomorphisms can be treated algebraically in a unified manner using the machinery of *Clifford algebras* and *spinors*; however, I find the more *ad hoc* geometric approach given here to be easier to visualise.

Remark 2.2.4. In the above discussion, we relied heavily on matching dimensions to ensure that various homomorphisms were in fact isogenies. There are some other exceptional homomorphisms in low dimension which are not isogenies due to mismatching dimensions, but are still of interest. For instance, there is a way to embed the six-dimensional space $D_2 = A_1 \times A_1 = C_1 \times B_1 = Sp_2(\mathbf{C}) \times SO_3(\mathbf{C})$ into the 21-dimensional space $C_3 = Sp_6(\mathbf{C})$, by letting $Sp_2(\mathbf{C})$ act on $\mathbf{C}^2$ and $SO_3(\mathbf{C})$ act on $\mathbf{C}^3$, so that $Sp_2(\mathbf{C}) \times SO_3(\mathbf{C})$ acts on the six-dimensional tensor product $\mathbf{C}^2 \otimes \mathbf{C}^3$ in the obvious manner; this preserves the tensor product of the symplectic form on $\mathbf{C}^2$ and the symmetric form on $\mathbf{C}^3$, which is a non-degenerate symplectic form on $\mathbf{C}^2 \otimes \mathbf{C}^3 \equiv \mathbf{C}^6$, giving the homomorphism (with the kernel once again being $\{(+1,+1),(-1,-1)\}$). These sorts of embeddings were useful in a recent paper of Breuillard, Green, Guralnick, and myself [**BrGrGuTa2010**], as they gave examples of semisimple groups that could be easily separated from other semisimple groups (such as $C_1 \times C_2$ inside C_3) due to their irreducible action on various natural vector spaces (i.e. they did not stabilise any non-trivial space).

Chapter 3

Combinatorics

3.1. The Szemerédi-Trotter theorem via the polynomial ham sandwich theorem

The *ham sandwich theorem* asserts that, given d bounded open sets $U_1, \ldots, U_d$ in $\mathbf{R}^d$, there exists a hyperplane $\{x \in \mathbf{R}^d : x \cdot v = c\}$ that bisects each of these sets U_i, in the sense that each of the two half-spaces $\{x \in \mathbf{R}^d : x \cdot v < c\}, \{x \in \mathbf{R}^d : x \cdot v > c\}$ on either side of the hyperplane captures exactly half of the volume of U_i. The shortest proof of this result proceeds by invoking the *Borsuk-Ulam theorem*.

A useful generalisation of the ham sandwich theorem is the *polynomial ham sandwich theorem*, which asserts that given m bounded open sets $U_1, \ldots, U_m$ in $\mathbf{R}^d$, there exists a hypersurface $\{x \in \mathbf{R}^d : Q(x) = 0\}$ of degree $O_d(m^{1/d})$ (thus $P : \mathbf{R}^d \to \mathbf{R}$ is a polynomial of degree[1] $O(m^{1/n})$ such that the two semi-algebraic sets $\{Q > 0\}$ and $\{Q < 0\}$ capture half the volume of each of the U_i. This theorem can be deduced from the Borsuk-Ulam theorem in the same manner that the ordinary ham sandwich theorem is (and can also be deduced directly from the ordinary ham sandwich theorem via the *Veronese embedding*).

The polynomial ham sandwich theorem is a theorem about continuous bodies (bounded open sets), but a simple limiting argument leads one to the following discrete analogue: given m *finite* sets $S_1, \ldots, S_m$ in $\mathbf{R}^d$, there exists a hypersurface $\{x \in \mathbf{R}^d : Q(x) = 0\}$ of degree $O_d(m^{1/d})$, such that each of the two semi-algebraic sets $\{Q > 0\}$ and $\{Q < 0\}$ contain at most half of the points on S_i (note that some of the points of S_i can certainly

[1] More precisely, the degree will be at most D, where D is the first positive integer for which $\binom{D+d}{d}$ exceeds m.

lie on the boundary $\{Q = 0\}$). This can be iterated to give a useful cell decomposition:

Proposition 3.1.1 (Cell decomposition)**.** *Let P be a finite set of points in $\mathbf{R}^d$, and let D be a positive integer. Then there exists a polynomial Q of degree at most D, and a decomposition*

$$\mathbf{R}^d = \{Q = 0\} \cup C_1 \cup \ldots \cup C_m$$

into the hypersurface $\{Q = 0\}$ and a collection $C_1, \ldots, C_m$ of cells bounded by $\{P = 0\}$, such that $m = O_d(D^d)$, and such that each cell C_i contains at most $O_d(|P|/D^d)$ points.

A proof of this decomposition is sketched in [**Ta2011d**, §3.9]. The cells in the argument are not necessarily connected (being instead formed by intersecting together a number of semi-algebraic sets such as $\{Q > 0\}$ and $\{Q < 0\}$), but it is a classical result[2] [**OlPe1949**], [**Mi1964**], [**Th1965**] that any degree D hypersurface $\{Q = 0\}$ divides $\mathbf{R}^d$ into $O_d(D^d)$ connected components, so one can easily assume that the cells are connected if desired.

Remark 3.1.2. By setting D as large as $O_d(|P|^{1/m})$, we obtain as a limiting case of the cell decomposition the fact that any finite set P of points in $\mathbf{R}^d$ can be captured by a hypersurface of degree $O_d(|P|^{1/m})$. This fact is in fact true over arbitrary fields (not just over $\mathbf{R}$), and can be proven by a simple linear algebra argument; see e.g. [**Ta2009b**, §1.1]. However, the cell decomposition is more flexible than this algebraic fact due to the ability to arbitrarily select the degree parameter D.

The cell decomposition can be viewed as a structural theorem for arbitrary large configurations of points in space, much as the *Szemerédi regularity lemma* [**Sz1978**] can be viewed as a structural theorem for arbitrary large dense graphs. Indeed, just as many problems in the theory of large dense graphs can be profitably attacked by first applying the regularity lemma and then inspecting the outcome, it now seems that many problems in combinatorial incidence geometry can be attacked by applying the cell decomposition (or a similar such decomposition), with a parameter D to be optimised later, to a relevant set of points, and seeing how the cells interact with each other and with the other objects in the configuration (lines, planes, circles, etc.). This strategy was spectacularly illustrated recently with Guth and Katz's use [**GuKa2010**] of the cell decomposition to resolve

[2]Actually, one does not need the full machinery of the results in the above cited papers - which control not just the number of components, but all the *Betti numbers* of the complement of $\{Q = 0\}$ - to get the bound on connected components; one can instead observe that every bounded connected component has a critical point where $\nabla Q = 0$, and one can control the number of these points by Bezout's theorem, after perturbing Q slightly to enforce genericity, and then count the unbounded components by an induction on dimension. See [**SoTa2011**, Appendix A].

the Erdös distinct distance problem (up to logarithmic factors), as discussed in [**Ta2011d**, §3.9].

In this section, I will record a simpler (but still illustrative) version of this method (that I learned from Nets Katz), which provides yet another proof of the *Szemerédi-Trotter* theorem in incidence geometry:

Theorem 3.1.3 (Szemerédi-Trotter theorem)**.** *Given a finite set of points P and a finite set of lines L in $\mathbf{R}^2$, the set of incidences $I(P,L) := \{(p,\ell) \in P \times L : p \in \ell\}$ has cardinality*

$$|I(P,L)| \ll |P|^{2/3}|L|^{2/3} + |P| + |L|.$$

This theorem has many short existing proofs, including one via crossing number inequalities (as discussed in [**Ta2008**, §1.10] or via a slightly different type of cell decomposition (as discussed in [**Ta2010b**, §1.6]). The proof given below is not that different, in particular, from the latter proof, but I believe it still serves as a good introduction to the polynomial method in combinatorial incidence geometry.

Let us begin with a trivial bound:

Lemma 3.1.4 (Trivial bound)**.** *For any finite set of points P and finite set of lines L, we have $|I(P,L)| \ll |P||L|^{1/2} + |L|$.*

The slickest way to prove this lemma is by the Cauchy-Schwarz inequality. If we let $\mu(\ell)$ be the number of points P incident to a given line ℓ, then we have

$$|I(P,L)| = \sum_{\ell \in L} \mu(\ell)$$

and hence by Cauchy-Schwarz

$$\sum_{\ell \in L} \mu(\ell)^2 \geq |I(P,L)|^2/|L|.$$

On the other hand, the left-hand side counts the number of triples $(p,p',\ell) \in P \times P \times L$ with $p,p' \in \ell$. Since two distinct points p,p' determine at most one line, one thus sees that the left-hand side is at most $|P|^2 + |I(P,L)|$, and the claim follows.

Now we return to the Szemerédi-Trotter theorem, and apply the cell decomposition with some parameter D. This gives a decomposition

$$\mathbf{R}^2 = \{Q = 0\} \cup C_1 \cup \ldots \cup C_m$$

into a curve $\{Q = 0\}$ of degree $O(D)$, and at most $O(D^2)$ cells $C_1, \ldots, C_m$, each of which contains $O(|P|/D^2)$ points. We can then decompose

$$|I(P,L)| = |I(P \cap \{Q=0\}, L)| + \sum_{i=1}^m |I(P \cap C_i, L)|.$$

By removing repeated factors, we may take Q to be square-free.

Let us first deal with the incidences coming from the cells C_i. Let L_i be the lines in L that pass through the i^{th} cell C_i. Clearly

$$|I(P \cap C_i, L)| = |I(P \cap C_i, L_i)|$$

and thus by the trivial bound

$$|I(P \cap C_i, L)| \ll |P \cap C_i||L_i|^{1/2} + |L_i| \ll \frac{|P|}{D^2}|L_i|^{1/2} + |L_i|.$$

Now we make a key observation (coming from *Bezout's theorem*: each line in ℓ can meet at most $O(D)$ cells C_i, because the cells C_i are bounded by a degree D curve $\{Q = 0\}$). Thus

$$\sum_{i=1}^{m} |L_i| \ll D|L|$$

and hence by Cauchy-Schwarz, we have

$$\sum_{i=1}^{m} |L_i|^{1/2} \ll D^{3/2}|L|^{1/2}.$$

Putting all this together, we see that

$$\sum_{i=1}^{m} |I(P \cap C_i, L)| \ll D^{-1/2}|P||L|^{1/2} + D|L|.$$

Now we turn to the incidences coming from the curve $\{Q = 0\}$. Applying Bezout's theorem again, we see that each line in L either lies in $\{Q = 0\}$, or meets $\{Q = 0\}$ in $O(D)$ points. The latter case contributes at most $O(D|L|)$ incidences, so now we restrict attention to lines that are completely contained in $\{Q = 0\}$. The points in the curve $\{Q = 0\}$ are of two types: smooth points (for which there is a unique tangent line to the curve $\{Q = 0\}$) and singular points (where Q and ∇Q both vanish). A smooth point can be incident to at most one line in $\{Q = 0\}$, and so this case contributes at most $|P|$ incidences. So we may restrict attention to the singular points. But by one last application of Bezout's theorem, each line in L can intersect the zero-dimensional set $\{Q = \nabla Q = 0\}$ in at most $O(D)$ points (note that each partial derivative of Q also has degree $O(D)$), giving another contribution of $O(D|L|)$ incidences. Putting everything together, we obtain

$$|I(P, L)| \ll D^{-1/2}|P||L|^{1/2} + D|L| + |P|$$

for any $D \geq 1$. An optimisation in D then gives the claim.

Remark 3.1.5. If one used the extreme case of the cell decomposition noted in Remark 3.1.2, one only obtains the trivial bound

$$|I(P, L)| \ll |P|^{1/2}|L| + |P|.$$

On the other hand, this bound holds over arbitrary fields k (not just over $\mathbf{R}$), and can be sharp in such cases (consider for instance the case when k is a finite field, P consists of all the points in k^2, and L consists of all the lines in k^2.)

3.2. A quantitative Kemperman theorem

In [**Ke1964**], Kemperman established the following result:

Theorem 3.2.1. *Let G be a compact connected group, with a Haar probability measure μ. Let A, B be compact subsets of G. Then*

$$\mu(AB) \geq \min(\mu(A) + \mu(B), 1).$$

Remark 3.2.2. The estimate is sharp, as can be seen by considering the case when G is a unit circle, and A, B are arcs; similarly if G is any compact connected group that projects onto the circle. The connectedness hypothesis is essential, as can be seen by considering what happens if A and B are a non-trivial open subgroup of G. For locally compact connected groups which are unimodular but not compact, there is an analogous statement, but with μ now a Haar measure instead of a Haar probability measure, and the right-hand side $\min(\mu(A) + \mu(B), 1)$ replaced simply by $\mu(A) + \mu(B)$. The case when G is a torus is due to Macbeath [**Ma1953**], and the case when G is a circle is due to Raikov [**Ra1939**]. The theorem is closely related to the Cauchy-Davenport inequality [**Ca1813**], [**Da1935**]; indeed, it is not difficult to use that inequality to establish the circle case, and the circle case can be used to deduce the torus case by considering increasingly dense circle subgroups of the torus (alternatively, one can also use Kneser's theorem [**Kn1953**]).

By inner regularity, the hypothesis that A, B are compact can be replaced with Borel measurability, so long as one then adds the additional hypothesis that $A + B$ is also Borel measurable.

A short proof of Kemperman's theorem was given by Ruzsa [**Ru1992**]. In this section, I wanted to record how this argument can be used to establish the following more "robust" version of Kemperman's theorem, which not only lower bounds AB, but gives many elements of AB some multiplicity:

Theorem 3.2.3. *Let G be a compact connected group, with a Haar probability measure μ. Let A, B be compact subsets of G. Then for any $0 \leq t \leq \min(\mu(A), \mu(B))$, one has*

$$\int_G \min(1_A * 1_B, t)\, d\mu \geq t \min(\mu(A) + \mu(B) - t, 1). \tag{3.1}$$

Indeed, Theorem 3.2.1 can be deduced from Theorem 3.2.3 by dividing (3.1) by t and then taking limits as $t \to 0$. The bound in (3.1) is sharp, as can again be seen by considering the case when A, B are arcs in a circle. The analogous claim for cyclic groups for prime order was established by Pollard [**Po1974**], and for general abelian groups by Green and Ruzsa [**GrRu2005**].

Let us now prove Theorem 3.2.3. It uses a submodularity argument related to some arguments of Hamidoune [**Ha2010**], [**Ta2012b**]. We fix B and t with $0 \leq t \leq \mu(B)$, and define the quantity

$$c(A) := \int_G \min(1_A * 1_B, t)\ d\mu - t(\mu(A) + \mu(B) - t).$$

for any compact set A. Our task is to establish that $c(A) \geq 0$ whenever $t \leq \mu(A) \leq 1 - \mu(B) + t$.

We first verify the extreme cases. If $\mu(A) = t$, then $1_A * 1_B \leq t$, and so $c(A) = 0$ in this case. At the other extreme, if $\mu(A) = 1 - \mu(B) + t$, then from the inclusion-exclusion principle we see that $1_A * 1_B \geq t$, and so again $c(A) = 0$ in this case (since $\int_G 1_A * 1_B = \mu(A)\mu(B) = t\mu(B)$).

To handle the intermediate regime when $\mu(A)$ lies between t and $1 - \mu(B) + t$, we rely on the *submodularity inequality*

$$c(A_1) + c(A_2) \geq c(A_1 \cap A_2) + c(A_1 \cup A_2) \tag{3.2}$$

for arbitrary compact A_1, A_2. This inequality comes from the obvious pointwise identity

$$1_{A_1} + 1_{A_2} = 1_{A_1 \cap A_2} + 1_{A_1 \cup A_2}$$

whence

$$1_{A_1} * 1_B + 1_{A_2} * 1_B = 1_{A_1 \cap A_2} * 1_B + 1_{A_1 \cup A_2} * 1_B$$

and thus (noting that the quantities on the left are closer to each other than the quantities on the right)

$$\min(1_{A_1} * 1_B, t) + \min(1_{A_2} * 1_B, t)$$

$$\geq \min(1_{A_1 \cap A_2} * 1_B, t) + \min(1_{A_1 \cup A_2} * 1_B, t)$$

at which point (3.2) follows by integrating over G and then using the inclusion-exclusion principle.

Now introduce the function

$$f(a) := \inf\{c(A) : \mu(A) = a\}$$

for $t \leq a \leq 1 - \mu(B) + t$. From the preceding discussion $f(a)$ vanishes at the endpoints $a = t, 1 - \mu(B) + t$; our task is to show that $f(a)$ is non-negative in the interior region $t < a < 1 - \mu(B) + t$. Suppose for contradiction that this was not the case. It is easy to see that f is continuous (indeed, it is even Lipschitz continuous), so there must be $t < a < 1 - \mu(B) + t$ at which

f is a local minimum and not locally constant. In particular, $0 < a < 1$. But for any A with $\mu(A) = a$, we have the translation-invariance

$$c(gA) = c(A) \tag{3.3}$$

for any $g \in G$, and hence by (3.2)

$$c(A) \geq \frac{1}{2}c(A \cap gA) + \frac{1}{2}c(A \cup gA).$$

Note that $\mu(A \cap gA)$ depends continuously on g, equals a when g is the identity, and has an average value of a^2. As G is connected, we thus see from the intermediate value theorem that for any $0 < \varepsilon < a - a^2$, we can find g such that $\mu(A \cap gA) = a - \varepsilon$, and thus by inclusion-exclusion $\mu(A \cup gA) = a + \varepsilon$. By definition of f, we thus have

$$c(A) \geq \frac{1}{2}f(a - \varepsilon) + \frac{1}{2}f(a + \varepsilon).$$

Taking infima in A (and noting that the hypotheses on ε are independent of A) we conclude that

$$f(a) \geq \frac{1}{2}f(a - \varepsilon) + \frac{1}{2}f(a + \varepsilon)$$

for all $0 < \varepsilon < a - a^2$. As f is a local minimum and ε is arbitrarily small, this implies that f is locally constant, a contradiction. This establishes Theorem 3.2.3.

We observe the following corollary:

Corollary 3.2.4. *Let G be a compact connected group, with a Haar probability measure μ. Let A, B, C be compact subsets of G, and let $\delta := \min(\mu(A), \mu(B), \mu(C))$. Then one has the pointwise estimate*

$$1_A * 1_B * 1_C \geq \frac{1}{4}(\mu(A) + \mu(B) + \mu(C) - 1)_+^2$$

if $\mu(A) + \mu(B) + \mu(C) - 1 \leq 2\delta$, and

$$1_A * 1_B * 1_C \geq \delta(\mu(A) + \mu(B) + \mu(C) - 1 - \delta)$$

if $\mu(A) + \mu(B) + \mu(C) - 1 \geq 2\delta$.

Once again, the bounds are completely sharp, as can be seen by computing $1_A * 1_B * 1_C$ when A, B, C are arcs of a circle. For groups G which are *quasirandom* (which means that they have no small-dimensional non-trivial representations, and are thus in some sense highly non-abelian), one can do much better than these bounds [**Go2008**]; thus, the abelian case is morally the worst case here, although it seems difficult to convert this intuition into a rigorous reduction.

Proof. By cyclic permutation we may take $\delta = \mu(C)$. For any

$$(\mu(A) + \mu(B) - 1)_+ \leq t \leq \min(\mu(A), \mu(B)),$$

we can bound

$$\begin{aligned} 1_A * 1_B * 1_C &\geq \min(1_A * 1_B, t) * 1_C \\ &\geq \int_G \min(1_A * 1_B, t)\ d\mu - t(1 - \mu(C)) \\ &\geq t(\mu(A) + \mu(B) - t) - t(1 - \mu(C)) \\ &= t \min(\mu(A) + \mu(B) + \mu(C) - 1 - t) \end{aligned}$$

where we used Theorem 3.2.3 to obtain the third line. Optimising in t, we obtain the claim. □

Chapter 4

Analysis

4.1. The Fredholm alternative

In one of my recent papers [**RoTa2011**], we needed to use the *Fredholm alternative* in functional analysis:

Theorem 4.1.1 (Fredholm alternative)**.** *Let X be a Banach space, let $T : X \to X$ be a* compact operator *(that is, a bounded linear operator that maps bounded sets to precompact sets), and let $\lambda \in \mathbf{C}$ be non-zero. Then exactly one of the following statements hold:*

(1) *(Eigenvalue) There is a non-trivial solution $x \in X$ to the equation $Tx = \lambda x$.*

(2) *(Bounded resolvent) The operator $T - \lambda$ has a bounded inverse $(T - \lambda)^{-1}$ on X.*

Among other things, the Fredholm alternative can be used to establish the *spectral theorem for compact operators*. A hypothesis such as compactness is necessary; the shift operator U on $\ell^2(\mathbf{Z})$, for instance, has no eigenfunctions, but $U - z$ is not invertible for any unit complex number z. The claim is also false when $\lambda = 0$; consider for instance the multiplication operator $Tf(n) := \frac{1}{n} f(n)$ on $\ell^2(\mathbf{N})$, which is compact and has no eigenvalue at zero, but is not invertible.

In this section we present a proof of the Fredholm alternative (first discovered by MacCleur-Hulland [**MaHu2008**] and by Uuye [**Uu2010**]) in the case of *approximable* operators, which are a special subclass of compact operators that are the limit of finite rank operators in the uniform topology.

Many Banach spaces (and in particular, all Hilbert spaces) have the *approximation property*[1] that implies (by a result of Grothendieck [**Gr1955**]) that all compact operators on that space are approximable. For instance, if X is a Hilbert space, then any compact operator is approximable, because any compact set can be approximated by a finite-dimensional subspace, and in a Hilbert space, the orthogonal projection operator to a subspace is always a contraction. In more general Banach spaces, finite-dimensional subspaces are still complemented, but the operator norm of the projection can be large. Indeed, there are examples of Banach spaces for which the approximation property fails; the first such examples were discovered by Enflo [**En1973**], and a subsequent paper by Alexander [**Al1974**] demonstrated the existence of compact operators in certain Banach spaces that are not approximable.

We also give two more traditional proofs of the Fredholm alternative, not requiring the operator to be approximable, which are based on the Riesz lemma and a continuity argument respectively.

4.1.1. First proof (approximable case only). In the finite-dimensional case, the Fredholm alternative is an immediate consequence of the *rank-nullity theorem*, and the finite rank case can be easily deduced from the finite dimensional case by some routine algebraic manipulation. The main difficulty in proving the alternative is to be able to take limits and deduce the approximable case from the finite rank case. The key idea of the proof is to use the approximable property to establish a lower bound on $T - \lambda I$ that is stable enough to allow one to take such limits.

Fix a non-zero λ. It is clear that T cannot have both an eigenvalue and bounded resolvent at λ, so now suppose that T has no eigenvalue at λ, thus $T - \lambda$ is injective. We claim that this implies a lower bound:

Lemma 4.1.2 (Lower bound)**.** *Let $\lambda \in \mathbf{C}$ be non-zero, and suppose that $T : X \to X$ be a compact operator that has no eigenvalue at λ. Then there exists $c > 0$ such that $\|(T - \lambda)x\| \geq c\|x\|$ for all $x \in X$.*

Proof. By homogeneity, it suffices to establish the claim for unit vectors x. Suppose this is not the case; then we can find a sequence of unit vectors x_n such that $(T-\lambda)x_n$ converges strongly to zero. Since λx_n has norm bounded away from zero (here we use the non-zero nature of λ), we conclude in particular that $y_n := Tx_n$ has norm bounded away from zero for sufficiently large n. By compactness of T, we may (after passing to a subsequence) assume that the y_n converge strongly to a limit y, which is thus also non-zero.

[1]The approximation property has many formulations; one of them is that the identity operator is the limit of a sequence of finite rank operators in the strong operator topology.

On the other hand, applying the bounded operator T to the strong convergence $(T-\lambda)x_n \to 0$ (and using the fact that T commutes with $T-\lambda$) we see that $(T-\lambda)y_n$ converges strongly to 0. Since y_n converges strongly to y, we conclude that $(T-\lambda)y = 0$, and thus we have an eigenvalue of T at λ, contradiction. □

Remark 4.1.3. Note that this argument is *ineffective* in that it provides no explicit value of c (and thus no explicit upper bound for the operator norm of the resolvent $(T-\lambda)^{-1}$). This is not surprising, given that the fact that T has no eigenvalue at λ is an open condition rather than a closed one, and so one does not expect bounds that utilise this condition to be uniform. (Indeed, the resolvent needs to blow up as one approaches the spectrum of T.)

From the lower bound, we see that to prove the bounded invertibility of $T-\lambda$, it will suffice to establish surjectivity. (Of course, we could have also obtained this reduction by using the *open mapping theorem*.) In other words, we need to establish that the range $\mathrm{Ran}(T-\lambda)$ of $T-\lambda$ is all of X.

Let us first deal with the easy case when T has *finite rank*, so that $\mathrm{Ran}(T)$ is some finite-dimension n. This implies that the kernel $\mathrm{Ker}(T)$ has codimension n, and we may thus split $X = \mathrm{Ker}(T) + Y$ for some n-dimensional space Y. The operator $T-\lambda$ is a non-zero multiple of the identity on $\mathrm{Ker}(T)$, and so $\mathrm{Ran}(T-\lambda)$ already contains $\mathrm{Ker}(T)$. On the other hand, the operator $T(T-\lambda)$ maps the n-dimensional space Y to the n-dimensional space $\mathrm{Ran}(T)$ injectively (since Y avoids $\mathrm{Ker}(T)$ and $T-\lambda$ is injective), and thus also surjectively (by the *rank-nullity theorem*). Thus $T(\mathrm{Ran}(T-\lambda))$ contains $\mathrm{Ran}(T)$, and thus (by the short exact sequence $0 \to \mathrm{Ker}(T) \to X \to \mathrm{Ran}(T) \to 0$) $\mathrm{Ran}(T-\lambda)$ is in fact all of X, as desired.

Finally, we deal with the case when T is approximable. The lower bound in Lemma 4.1.2 is stable, and will extend to the finite rank operators S_n for n large enough (after reducing c slightly). By the preceding discussion for the finite rank case, we see that $\mathrm{Ran}(S_n-\lambda)$ is all of X. Using Lemma 4.1.2 for S_n, and the convergence of S_n to T in the operator norm topology, we conclude that $\mathrm{Ran}(T-\lambda)$ is dense in X. On the other hand, we observe that the space $\mathrm{Ran}(T-\lambda)$ is necessarily closed, for if $(T-\lambda)x_n$ converges to a limit y, then (by Lemma 4.1.2 and the assumption that X is Banach) x_n will also converge to some limit x, and so $y = (T-\lambda)x$. As $\mathrm{Ran}(T-\lambda)$ is now both dense and closed, it must be all of X, and the claim follows.

4.1.2. Second proof. We now give the standard proof of the Fredholm alternative based on the *Riesz lemma*:

Lemma 4.1.4 (Riesz lemma)**.** *If Y is a proper closed subspace of a Banach space X, and $\varepsilon > 0$, then there exists a unit vector x whose distance* $\mathrm{dist}(x, Y)$ *to Y is at least $1 - \varepsilon$.*

Proof. By the *Hahn-Banach theorem*, one can find a non-trivial linear functional $\phi : X \to \mathbf{C}$ on X which vanishes on Y. By definition of the operator norm $\|\phi\|_{\mathrm{op}}$ of ϕ, one can find a unit vector x such that $|\phi(x)| \geq (1-\varepsilon)\|\phi\|_{\mathrm{op}}$. The claim follows. □

The strategy here is not to use finite rank approximations (as they are no longer available), but instead to try to contradict the compactness of T by exhibiting a bounded set whose image under T is not *totally bounded*.

Let $T : X \to X$ be a compact operator on a Banach space, and let λ be a non-zero complex number such that T has no eigenvalue at λ. As in the first proof, we have the lower bound from Lemma 4.1.2, and we know that $\mathrm{Ran}(T - \lambda)$ is a closed subspace of X; in particular, the map $T - \lambda$ is a Banach space isomorphism from X to $\mathrm{Ran}(T - \lambda)$. Our objective is again to show that $\mathrm{Ran}(T - \lambda)$ is all of X.

Suppose for contradiction that $\mathrm{Ran}(T-\lambda)$ is a proper closed subspace of X. Applying the Banach space isomorphism $T - \lambda$ repeatedly, we conclude that for every natural number m, the space $V_{m+1} := \mathrm{Ran}((T - \lambda)^{m+1})$ is a proper closed subspace of $V_m := \mathrm{Ran}((T-\lambda)^m)$. From the Riesz lemma, we may thus find unit vectors x_m in V_m for $m = 0, 1, 2, \ldots$ whose distance to V_{m+1} is at least $1/2$ (say).

Now suppose that $n > m \geq 0$. By construction, $x_n, (T-\lambda)x_n, (T-\lambda)x_m$ all lie in V_{m+1}, and thus $Tx_n - Tx_m \in \lambda x_m + V_{m+1}$. Since x_m lies at a distance at least $1/2$ from V_{m+1}, we conclude the separation proeprty

$$\|Tx_n - Tx_m\| \geq \frac{|\lambda|}{2}.$$

But this implies that the sequence $\{Tx_n : n \in \mathbf{N}\}$ is not totally bounded, contradicting the compactness of T.

4.1.3. Third proof. Now we give another textbook proof of the Fredholm alternative, based on Fredholm index theory. The basic idea is to observe that the Fredholm alternative is easy when λ is large enough (and specifically, when $|\lambda| > \|T\|_{\mathrm{op}}$), as one can then invert $T-\lambda$ using *Neumann series*. One can then attempt to continously perturb λ from large values to small values, using stability results (such as Lemma 4.1.2) to ensure that invertibility does not suddenly get destroyed during this process. Unfortunately, there is an obstruction to this strategy, which is that during the perturbation process, λ may pass through an eigenvalue of T. To get around this, we will need to abandon the hypothesis that T has no eigenvalue at λ, and work

in the more general setting in which $\mathrm{Ker}(T-\lambda)$ is allowed to be non-trivial. This leads to a lengthier proof, but one which lays the foundation for much of *Fredholm theory* (which is more powerful than the Fredholm alternative alone).

Fortunately, we still have analogues of much of the above theory in this setting:

Proposition 4.1.5. *Let $\lambda \in \mathbf{C}$ be non-zero, and let $T : X \to X$ be a compact operator on a Banach space X. Then the following statements hold;*

(1) *(Finite multiplicity)* $\mathrm{Ker}(T-\lambda)$ *is finite-dimensional.*

(2) *(Lower bound) There exists $c > 0$ such that $\|Tx\| \geq c\,\mathrm{dist}(x, \mathrm{Ker}(T-\lambda))$ for all $x \in X$.*

(3) *(Closure)* $\mathrm{Ran}(T-\lambda)$ *is a closed subspace of X.*

(4) *(Finite comultiplicity)* $\mathrm{Ran}(T-\lambda)$ *has finite codimension in X.*

Proof. We begin with finite multiplicity. Suppose for contradiction that $\mathrm{Ker}(T-\lambda)$ was infinite dimensional, then it must contain an infinite nested sequence $\{0\} = V_0 \subsetneq V_1 \subsetneq V_2 \subsetneq \dots$ of finite-dimensional (and thus closed) subspaces. Applying the Riesz lemma, we may find for each $n = 1, 2, \dots$, a unit vector $x_n \in V_n$ of distance at least $1/2$ from V_{n-1}. Since $Tx_n = \lambda x_n$, we see that the sequence $\{Tx_n : n = 1, 2, \dots\}$ is then $|\lambda|/2$-separated and thus not totally bounded, contradicting the compactness of T.

The lower bound follows from the argument used to prove Lemma 4.1.2 after quotienting out the finite-dimensional space $\mathrm{Ker}(T-\lambda)$, and the closure assertion follows from the lower bound (again after quotienting out the kernel) as before.

Finally, we establish finite comultiplicity. Suppose for contradiction that the closed subspace $\mathrm{Ran}(T-\lambda)$ had infinite codimension, then by properties of $T-\lambda$ already established, we see that $\mathrm{Ran}((T-\lambda)^{m+1})$ is closed and has infinite codimension in $\mathrm{Ran}((T-\lambda)^m)$ for each m. One can then argue as in the second proof to contradict total boundedness as before. □

Remark 4.1.6. The above arguments also work if λ is replaced by an invertible linear operator on X, or more generally by a *Fredholm operator*.

We can now define the *index* $\mathrm{ind}(T-\lambda)$ to be the dimension of the kernel of $T-\lambda$, minus the codimension of the range. To establish the Fredholm alternative, it suffices to show that $\mathrm{ind}(T-\lambda) = 0$ for all λ, as this implies surjectivity of $T-\lambda$ whenever there is no eigenvalue. Note that Note that when λ is sufficiently large, and in particular when $|\lambda| > \|T\|_{\mathrm{op}}$, then $T-\lambda$ is invertible by Neumann series and so one already has index zero in this case.

To finish the proof, it suffices by the discrete nature of the index function (which takes values in the integers) to establish continuity of the index:

Lemma 4.1.7 (Continuity of index)**.** *Let $T : X \to X$ be a compact operator on a Banach space. Then the function $\lambda \mapsto \mathrm{ind}(T - \lambda)$ is continuous from $\mathbf{C}\backslash\{0\}$ to $\mathbf{Z}$.*

Proof. Let λ be non-zero. Our task is to show that

$$\mathrm{ind}(T - \lambda') = \mathrm{ind}(T - \lambda)$$

for all λ' sufficiently close to λ.

In the model case when $T - \lambda$ is invertible (and thus has index zero), the claim is easy, because $(T - \lambda')(T - \lambda)^{-1} = 1 + (\lambda - \lambda')(T - \lambda)^{-1}$ can be inverted by Neumann series for λ' close enough to λ, giving rise to the invertibility of $T - \lambda$.

Now we handle the general case. As every finite dimensional space is complemented, we can split $X = \mathrm{Ker}(T - \lambda) + V$ for some closed subspace V of X, and similarly split $X = \mathrm{Ran}(T - \lambda) + W$ for some finite-dimensional subspace W of X with dimension $\mathrm{codim}\,\mathrm{Ran}(T - \lambda)$.

From the lower bound we see that $T - \lambda$ is a Banach space isomorphism from V to $\mathrm{Ran}(T - \lambda)$. For λ' close to λ, we thus see that $(T - \lambda')(V)$ is close to $\mathrm{Ran}(T - \lambda)$, in the sense that one can map the latter space to the former by a small perturbation of the identity (in the operator norm). Since W complements $\mathrm{Ran}(T - \lambda)$, it also complements $(T - \lambda')(V)$ for λ' sufficiently close to λ. (To see this, observe that the composition of the obvious maps

$$X \mapsto W \times \mathrm{Ran}(T - \lambda) \to W \times V \to W \times (T - \lambda')(V) \to X$$

is a small perturbation of the identity map and is thus invertible for λ' close to λ.)

Let $\pi : X \to W$ be the projection onto W with kernel $(T - \lambda')(V)$. Then $\pi(T - \lambda')$ maps the finite-dimensional space $\mathrm{Ker}(T - \lambda)$ to the finite-dimensional space W. By the *rank-nullity theorem*, this map has index equal to $\dim \mathrm{Ker}(T - \lambda) - \dim(W) = \mathrm{ind}(T - \lambda)$. Gluing this with the Banach space isomorphism $T - \lambda' : V \to \mathrm{Ran}(T - \lambda')$, we see that $T - \lambda'$ also has index $\mathrm{ind}(T - \lambda)$, as desired. □

Remark 4.1.8. Again, this result extends to more general Fredholm operators, with the result being that the index of a Fredholm operator is stable with respect to continuous deformations in the operator norm topology.

4.2. The inverse function theorem for everywhere differentiable functions

The classical *inverse function theorem* reads as follows:

Theorem 4.2.1 (C^1 inverse function theorem)**.** *Let $\Omega \subset \mathbf{R}^n$ be an open set, and let $f : \Omega \to \mathbf{R}^n$ be an continuously differentiable function, such that for every $x_0 \in \Omega$, the derivative map $Df(x_0) : \mathbf{R}^n \to \mathbf{R}^n$ is invertible. Then f is a local homeomorphism; thus, for every $x_0 \in \Omega$, there exists an open neighbourhood U of x_0 and an open neighbourhood V of $f(x_0)$ such that f is a homeomorphism from U to V.*

It is also not difficult to show by inverting the Taylor expansion

$$f(x) = f(x_0) + Df(x_0)(x - x_0) + o(\|x - x_0\|)$$

that at each x_0, the local inverses $f^{-1} : V \to U$ are also differentiable at $f(x_0)$ with derivative

$$Df^{-1}(f(x_0)) = Df(x_0)^{-1}. \tag{4.1}$$

The textbook proof of the inverse function theorem proceeds by an application of the *contraction mapping theorem*. Indeed, one may normalise $x_0 = f(x_0) = 0$ and $Df(0)$ to be the identity map; continuity of Df then shows that $Df(x)$ is close to the identity for small x, which may be used (in conjunction with the fundamental theorem of calculus) to make $x \mapsto x - f(x) + y$ a contraction on a small ball around the origin for small y, at which point the contraction mapping theorem readily finishes off the problem.

Less well known is the fact that the hypothesis of continuous differentiability may be relaxed to just everywhere differentiability:

Theorem 4.2.2 (Everywhere differentiable inverse function theorem)**.** *Let $\Omega \subset \mathbf{R}^n$ be an open set, and let $f : \Omega \to \mathbf{R}^n$ be an everywhere differentiable function, such that for every $x_0 \in \Omega$, the derivative map $Df(x_0) : \mathbf{R}^n \to \mathbf{R}^n$ is invertible. Then f is a local homeomorphism; thus, for every $x_0 \in \Omega$, there exists an open neighbourhood U of x_0 and an open neighbourhood V of $f(x_0)$ such that f is a homeomorphism from U to V.*

As before, one can recover the differentiability of the local inverses, with the derivative of the inverse given by the usual formula (4.1).

This result implicitly follows from the more general results of Cernavskii [**Ce1964**] about the structure of finite-to-one open and closed maps, however the arguments there are somewhat complicated (and subsequent proofs of those results, such as the one in [**Va1966**], use some powerful tools from algebraic topology, such as *dimension theory*). There is however a more elementary proof of Saint Raymond [**Ra2002**] that was pointed out to me by Julien Melleray. It only uses basic point-set topology (for instance, the concept of a connected component) and the basic topological and geometric

structure of Euclidean space (in particular relying primarily on local compactness, local connectedness, and local convexity). I decided to present (an arrangement of) Saint Raymond's proof here.

To obtain a local homeomorphism near x_0, there are basically two things to show: local surjectivity near x_0 (thus, for y near $f(x_0)$, one can solve $f(x) = y$ for some x near x_0) and local injectivity near x_0 (thus, for distinct x_1, x_2 near $f(x_0)$, $f(x_1)$ is not equal to $f(x_2)$). Local surjectivity is relatively easy; basically, the standard proof of the inverse function theorem works here, after replacing the contraction mapping theorem (which is no longer available due to the possibly discontinuous nature of Df) with the *Brouwer fixed point theorem instead* (or one could also use *degree theory*, which is more or less an equivalent approach). The difficulty is local injectivity - one needs to preclude the existence of nearby points x_1, x_2 with $f(x_1) = f(x_2) = y$; note that in contrast to the contraction mapping theorem that provides both existence and uniqueness of fixed points, the Brouwer fixed point theorem only gives existence and not uniqueness.

In one dimension $n = 1$ one can proceed by using *Rolle's theorem*. Indeed, as one traverses the interval from x_1 to x_2, one must encounter some intermediate point x_* which maximises the quantity $|f(x_*) - y|$, and which is thus instantaneously non-increasing both to the left and to the right of x_*. But, by hypothesis, $f'(x_*)$ is non-zero, and this easily leads to a contradiction.

Saint Raymond's argument for the higher dimensional case proceeds in a broadly similar way. Starting with two nearby points x_1, x_2 with $f(x_1) = f(x_2) = y$, one finds a point x_* which "locally extremises" $\|f(x_*) - y\|$ in the following sense: $\|f(x_*) - y\|$ is equal to some $r_* > 0$, but x_* is adherent to at least two distinct connected components U_1, U_2 of the set $U = \{x : \|f(x) - y\| < r_*\}$. (This is an oversimplification, as one has to restrict the available points x in U to a suitably small compact set, but let us ignore this technicality for now.) Note from the non-degenerate nature of $Df(x_*)$ that x_* was already adherent to U; the point is that x_* "disconnects" U in some sense. Very roughly speaking, the way such a critical point x_* is found is to look at the sets $\{x : \|f(x) - y\| \leq r\}$ as r shrinks from a large initial value down to zero, and one finds the first value of r_* below which this set disconnects x_1 from x_2. (Morally, one is performing some sort of *Morse theory* here on the function $x \mapsto \|f(x) - y\|$, though this function does not have anywhere near enough regularity for classical Morse theory to apply.)

The point x_* is mapped to a point $f(x_*)$ on the boundary $\partial B(y, r_*)$ of the ball $B(y, r_*)$, while the components U_1, U_2 are mapped to the interior of this ball. By using a continuity argument, one can show (again very roughly

speaking) that $f(U_1)$ must contain a "hemispherical" neighbourhood $\{z \in B(y, r_*) : \|z - f(x_*)\| < \kappa\}$ of $f(x_*)$ inside $B(y, r_*)$, and similarly for $f(U_2)$. But then from differentiability of f at x_*, one can then show that U_1 and U_2 overlap near x_*, giving a contradiction.

We now give the rigorous argument. Fix $x_0 \in \Omega$. By a translation, we may assume $x_0 = f(x_0) = 0$; by a further linear change of variables, we may also assume $Df(0)$ (which by hypothesis is non-singular) to be the identity map. By differentiability, we have

$$f(x) = x + o(\|x\|)$$

as $x \to 0$. In particular, there exists a ball $B(0, r_0)$ in Ω such that

$$\|f(x) - x\| < \frac{1}{2}\|x\|$$

for all $x \in B(0, r_0)$; by rescaling we may take $r_0 = 1$, thus

$$\|f(x) - x\| < \frac{1}{2}\|x\| \text{ whenever } \|x\| \leq 1. \tag{4.2}$$

Among other things, this gives a uniform lower bound

$$\|f(x)\| > \frac{1}{2} \tag{4.3}$$

for all $x \in \partial B(0, 1)$, and a uniform upper bound

$$\|f(x)\| < \frac{1}{10} \tag{4.4}$$

for all $x \in \partial B(0, \frac{1}{20})$; thus f maps $B(0, \frac{1}{20})$ to $B(0, \frac{1}{10})$.

Proposition 4.2.3 (Local surjectivity)**.** *For any $0 < r < 1$, $f(B(0, r))$ contains $B(0, r/2)$.*

Proof. Let $y \in B(0, r/2)$. From (4.2), we see that the map $f : \partial B(0, r) \to f(\partial B(0, r))$ avoids y, and has *degree* 1 around y; contracting $\partial B(0, r)$ to a point, we conclude that $f(x) = y$ for some $x \in B(0, r)$, yielding the claim.

Alternatively, one may proceed by invoking the *Brouwer fixed point theorem*, noting that the map $x \mapsto x - f(x) + y$ is continuous and maps the closed ball $\overline{B(0, r)}$ to the open ball $B(0, r)$ by (4.2), and has a fixed point precisely when $f(x) = y$.

A third argument (avoiding the use of degree theory or the Brouwer fixed point theorem, but requiring one to replace $B(0, r/2)$ with the slightly smaller ball $B(0, r/3)$) is as follows: let $x \in \overline{B(0, r)}$ minimise $\|f(x) - y\|$. From (4.2) and the hypothesis $y \in B(0, r/3)$ we see that x lies in the interior $B(0, r)$. If the minimum is zero, then we have found a solution to $f(x) = y$ as required; if not, then we have a stationary point of $x \mapsto \|f(x) - y\|$, which implies that $Df(x)$ is degenerate, a contradiction. (One can recover the full

ball $B(0,r/2)$ by tweaking the expression $\|f(x)-y\|$ to be minimised in a suitable fashion; we leave this as an exercise for the interested reader.) □

Corollary 4.2.4. *f is an* open map*: the image of any open set is open.*

Proof. It suffices to show that for every $x \in \Omega$, the image of any open neighbourhood of x is an open neighbourhood of $f(x)$. Proposition 4.2.3 handles the case $x = 0$; the general case follows by renormalising. □

Suppose we could show that f is injective on $B(0, \frac{1}{20})$. By Corollary 4.2.4, the inverse map $f^{-1} : f(B(0, \frac{1}{20})) \to B(0, \frac{1}{20})$ is also continuous. Thus f is a homeomorphism from $B(0, \frac{1}{20})$ to $f(B(0, \frac{1}{20}))$, which are both neighbourhoods of 0 by Proposition 4.2.3; giving the claim.

It remains to establish injectivity. Suppose for sake of contradiction that this was not the case. Then there exists $x_1, x_2 \in B(0, \frac{1}{20})$ and $y \in B(0, \frac{1}{10})$ such that

$$y = f(x_1) = f(x_2).$$

For every radius $r \geq 0$, the set

$$K_r := \{x \in \Omega : \|f(x) - y\| \leq r\}$$

is closed and contains both x_1 and x_2. Let K_r^1 denote the connected component of K_r that contains x_1. Since K_r is non-decreasing in r, K_r^1 is non-decreasing also.

Now let us study the behaviour of K_r^1 as r ranges from 0 to $\frac{4}{10}$. The two extreme cases are easy to analyse:

Lemma 4.2.5. $K_0^1 = \{x_1\}$.

Proof. Since $Df(x_1)$ is non-singular, we see from differentiability that $f(x) \neq f(x_1)$ for all $x \neq x_1$ sufficiently close to x_1. Thus x_1 is an isolated point of K_0, and the claim follows. □

Lemma 4.2.6. *We have $B(0, \frac{1}{20}) \subset K_r^1 \subset B(0,1)$ for all $\frac{2}{10} \leq r \leq \frac{4}{10}$. In particular, K_r^1 is compact for all $0 \leq r \leq \frac{4}{10}$, and contains x_2 for $\frac{2}{10} \leq r \leq \frac{4}{10}$.*

Proof. Since $f(B(0, \frac{1}{20})) \subset B(f(0), \frac{1}{10}) \subset \overline{B(y,r)}$, we see that $B(0, \frac{1}{20}) \subset K_r$; since $B(0, \frac{1}{20})$ is connected and contains x_1, we conclude that $B(0, \frac{1}{20}) \subset K_r^1$.

Next, if $x \in \partial B(0,1)$, then by (4.3) we have $f(x) \notin B(0, \frac{1}{2})$, and hence $f(x) \notin \overline{B(y,r)}$. Thus K_r is disjoint from the sphere $\partial B(0,1)$. Since x_1 lies in the interior of this sphere we thus have $K_r^1 \subset B(0,1)$ as required. □

Next, we show that the K_r^1 increase continuously in r:

Lemma 4.2.7. *If $0 \leq r < \frac{1}{20}$ and $\varepsilon > 0$, then for $r < r' < \frac{1}{20}$ sufficiently close to r, $K^1_{r'}$ is contained in an ε-neighbourhood of K^1_r.*

Proof. By the finite intersection property, it suffices to show that $\bigcap_{r'>r} K^1_{r'} = K^1_r$. Suppose for contradiction that there is a point x outside of K^1_r that lies in $K^1_{r'}$ for all $r' > r$. Then x lies in $K_{r'}$ for all $r' > r$, and hence lies in $K_r \cap B(0,1)$. As x and x_1 lie in different connected components of the compact set $K_r \cap \overline{B(0,1)}$ (recall that K_r is disjoint from $\partial B(0,1)$), there must be a partition of $K_r \cap \overline{B(0,1)}$ into two disjoint closed sets F, G that separate x from x_1 (for otherwise the only clopen sets in $K_r \cap \overline{B(0,1)}$ that contain x_1 would also contain x, and their intersection would then be a connected subset of $K_r \cap \overline{B(0,1)}$ that contains both x_1 and x, contradicting the fact that x lies outside K^1_r). By normality, we may find open neighbourhoods U, V of F, G that are disjoint. For all x on the boundary ∂U, one has $\|f(x) - y\| > r$ for all $x \in \partial U$. As ∂U is compact and f is continuous, we thus have $\|f(x) - y\| > r'$ for all $x \in \partial U$ if r' is sufficiently close to r. This makes $U \cap K_{r'}$ clopen in $K_{r'}$, and so x cannot lie in $K^1_{r'}$, giving the desired contradiction. □

Observe that K^1_r contains x_2 for $r \geq \frac{2}{10}$, but does not contain x_2 for $r = 0$. By the monotonicity of the K^1_r and least upper bound principle, there must therefore exist a critical $0 \leq r_* \leq \frac{2}{10}$ such that K^1_r contains x_2 for all $r > r_*$, but does not contain x_2 for $r < r_*$. From Lemma 4.2.7 we see that $K^1_{r_*}$ must also contain x_2. In particular, by Lemma 4.2.5, $r_* > 0$.

We now analyse the critical set $K^1_{r_*}$. By construction, this set is connected, compact, contains both x_1 and x_2, contained in $B(0,1)$, and one has $\|f(x) - y\| \leq r_*$ for all $x \in K^1_{r_*}$.

Lemma 4.2.8. *The set $U := \{x \in K^1_{r_*} : \|f(x) - y\| < r_*\}$ is open and disconnected.*

Proof. The openness is clear from the continuity of f (and the local connectedness of $\mathbf{R}^n$). Now we show disconnectedness. Being an open subset of $\mathbf{R}^n$, connectedness is equivalent to path connectedness, and x_1 and x_2 both lie in U, so it suffices to show that x_1 and x_2 cannot be joined by a path γ in U. But if such a path γ existed, then by compactness of γ and continuity of f, one would have $\gamma \subset K_r$ for some $r < r_*$. This would imply that $x_2 \in K^1_r$, contradicting the minimal nature of r_*, and the claim follows. □

Lemma 4.2.9. *U has at most finitely many connected components.*

Proof. Let U_1 be a connected component of U; then $f(U_1)$ is non-empty and contained in $B(y, r_*)$. As U is open, U_1 is also open, and thus by Corollary 4.2.4, $f(U_1)$ is open also.

We claim that $f(U_1)$ is in fact all of $B(y, r_*)$. Suppose this were not the case. As $B(y, r_*)$ is connected, this would imply that $f(U_1)$ is not closed in $B(y, r_*)$; thus there is an element z of $B(y, r_*)$ which is adherent to $f(U_1)$, but does not lie in $f(U_1)$. Thus one may find a sequence x_n in U_1 with $f(x_n)$ converging to z. By compactness of $K^1_{r_*}$ (which contains U_1), we may pass to a subsequence and assume that x_n converges to a limit x in $K^1_{r_*}$; then $f(x) = z$. By continuity, there is thus a ball B centred at x that is mapped to $B(y, r)$ for some $r < r_*$; this implies that B lies in K_{r_*} and hence in $K^1_{r_*}$ (since $x \in K^1_{r_*}$) and thence in U (since r is strictly less than r_*). As x is adherent to U_1 and B is connected, we conclude that B lies in U_1. In particular x lies in U_1 and so $z = f(x)$ lies in $f(U_1)$, a contradiction.

As $f(U_1)$ is equal to $B(y, r_*)$, we thus see that U_1 contains an element of $f^{-1}(\{y\})$. However, each element x of $f^{-1}(\{y\})$ must be isolated since $Df(x)$ is non-singular. By compactness of $K^1_{r_*}$, the set $K^1_{r_*}$ (and hence U) thus contains at most finitely many elements of $f^{-1}(\{y\})$, and so there are finitely many components as claimed. □

Lemma 4.2.10. *Every point in $K^1_{r_*}$ is adherent to U (i.e. $\overline{U} = K^1_{r_*}$).*

Proof. If $x \in K^1_{r_*}$, then $\|f(x) - y\| \leq r_*$. If $\|f(x) - y\| < r_*$ then $x \in U$ and we are done, so we may assume $\|f(x) - y\| = r_*$. By differentiability, one has

$$f(x') = f(x) + Df(x)(x' - x) + o(\|x' - x\|)$$

for all x' sufficiently close to x. If we choose x' to lie on a ray emenating from x such that $Df(x)(x' - x)$ lies on a ray pointing towards y from $f(x)$ (this is possible as $Df(x)$ is non-singular), we conclude that for all x' sufficiently close to x on this ray, $\|f(x') - y\| < r_*$. Thus all such points x' lie in K_{r_*}; since x lies in $K^1_{r_*}$ and the ray is locally connected, we see that all such points x' in fact lie in $K^1_{r_*}$ and thence in U. The claim follows. □

Corollary 4.2.11. *There exists a point $x_* \in K^1_{r_*}$ with $\|f(x_*) - y\| = r_*$ (i.e. x_* lies outside U) which is adherent to at least two connected components of U.*

Proof. Suppose this were not the case, then the closures of all the connected components of U would be disjoint. (Note that an element of one connected component of U cannot lie in the closure of another component.) By Lemma 4.2.10, these closures would form a partition of $K^1_{r_*}$ by closed sets. By Lemma 4.2.8, there are at least two such closed sets, each of which is non-empty; by Lemma 4.2.9, the number of such closed sets is finite. But this contradicts the connectedness of $K^1_{r_*}$. □

Next, we prove

Proposition 4.2.12. *Let $x_* \in K^1_{r_*}$ be such that $\|f(x_*) - y\| = r_*$, and suppose that x is adherent to a connected component U_1 of U. Let ω be the vector such that*

$$Df(x_*)\omega = y - f(x_*) \tag{4.5}$$

(this vector exists and is non-zero since $Df(x_*)$ is non-singular). Then U_1 contains an open ray of the form $\{x_* + t\omega : 0 < t < \varepsilon\}$ for some $\varepsilon > 0$.

This together with Corollary 4.2.11 gives the desired contradiction, since one cannot have two distinct components U_1, U_2 both contain a ray from x_* in the direction ω.

Proof. As f is differentiable at x_*, we have

$$f(x_* + t\omega) = f(x_*) + Df(x_*)t\omega + o(|t|)$$

for all sufficiently small t; we rearrange this using (4.5) as

$$f(x_* + t\omega) - y = (1-t)(f(x_*) - y) + o(|t|).$$

In particular, $f(x_* + t\omega) \in B(y, r_*)$ for all sufficiently small positive t. This shows that all sufficiently small open rays $\{x_* + t\omega : 0 < t < \varepsilon\}$ lie in K_{r_*}, hence in $K^1_{r_*}$ (since $x_* \in K^1_{r_*}$), and hence in U. In fact, the same argument shows that there is a cone

$$\{x_* + t\omega' : 0 < t < \varepsilon; \|\omega' - \omega\| \leq \varepsilon\} \tag{4.6}$$

that will lie in U if ε is small enough. As this cone is connected, it thus suffices to show that U_1 intersects this cone.

Let $\delta > 0$ be a small radius to be chosen later. As $Df(x_*)$ is non-singular, we see if δ is small enough that $f(x) \neq f(x_*)$ whenever $\|x - x_*\| = \delta$. By continuity, we may thus find $\kappa > 0$ such that $\|f(x) - f(x_*)\| > \kappa$ whenever $\|x - x_*\| = \delta$.

Consider the set

$$U' := \{x \in U_1 : \|x - x_*\| \leq \delta; \|f(x) - f(x_*)\| < \kappa\}.$$

As x_* is adherent to U_1, U' is non-empty. By construction of κ, we see that we also have

$$U' := \{x \in U_1 : \|x - x_*\| < \delta; \|f(x) - f(x_*)\| < \kappa\}$$

and so U' is open. By Corollary 4.2.4, $f(U')$ is then also non-empty and open. By construction, $f(U')$ also lies in the set

$$D := \{z \in B(y, r_*) : \|z - f(x_*)\| < \kappa\}.$$

We claim that $f(U')$ is in fact all of D. The proof will be a variant of the proof of Lemma 4.2.9. Suppose this were not the case. As D is connected, this implies that there is an element z of D which is adherent to $f(U')$, but

does not lie in $f(U')$. Thus one may find a sequence x_n in U' with $f(x_n)$ converging to z. By compactness of $K^1_{r_*}$ (which contains U'), we may pass to a subsequence and assume that x_n converges to a limit x in $K^1_{r_*}$; then $f(x) = z$. By continuity, there is thus a ball B centred at x contained in $B(x_*, \delta)$ that is mapped to $B(y, r) \cap D$ for some $r < r_*$; this implies that B lies in K_{r_*} and hence in $K^1_{r_*}$ (since $x \in K^1_{r_*}$) and thence in U (since r is strictly less than r_*). As x is adherent to U_1 and B is connected, we conclude that B lies in U_1 and thence in U'. In particular x lies in U' and so $z = f(x)$ lies in $f(U')$, a contradiction.

As $f(U') = D$, we may thus find a sequence $t_n > 0$ converging to zero, and a sequence $x_n \in U'$, such that

$$f(x_n) = f(x_*) + t_n(y - f(x_*)).$$

However, if δ is small enough, we have $\|f(x_n) - f(x_*)\|$ comparable to $\|x_n - x_*\|$ (cf. (4.2)), and so x_n converges to x_*. By Taylor expansion, we then have

$$f(x_n) = f(x_*) + Df(x_*)(x_n - x_*) + o(\|x_n - x_*\|)$$

and thus

$$(Df(x_*) + o(1))(x_n - x_*) = t_n Df(x_*)\omega$$

for some matrix-valued error $o(1)$. Since $Df(x_*)$ is invertible, this implies that

$$x_n - x_* = t_n(1 + o(1))\omega = t_n\omega + o(t_n).$$

In particular, x_n lies in the cone (4.6) for n large enough, and the claim follows. □

4.3. Stein's interpolation theorem

One of Eli Stein's very first results that is still used extremely widely today, is his interpolation theorem [**St1956**] (and its refinement, the Fefferman-Stein interpolation theorem [**FeSt1972**]). This is a deceptively innocuous, yet remarkably powerful, generalisation of the classic *Riesz-Thorin interpolation theorem* (see e.g. [**Ta2010**, Theorem 1.11.7]) which uses methods from complex analysis (and in particular, the *Lindelöf theorem* or the *Phragmén-Lindelöf principle*) to show that if a linear operator $T : L^{p_0}(X) + L^{p_1}(X) \to L^{q_0}(Y) + L^{q_1}(Y)$ from one (σ-finite) measure space $X = (X, \mathcal{X}, \mu)$ to another $Y = (Y, \mathcal{Y}, \nu)$ obeyed the estimates

$$\|Tf\|_{L^{q_0}(Y)} \leq B_0\|f\|_{L^{p_0}(X)} \tag{4.7}$$

for all $f \in L^{p_0}(X)$ and

$$\|Tf\|_{L^{q_1}(Y)} \leq B_1\|f\|_{L^{p_1}(X)} \tag{4.8}$$

for all $f \in L^{p_1}(X)$, where $1 \leq p_0, p_1, q_0, q_1 \leq \infty$ and $B_0, B_1 > 0$, then one automatically also has the interpolated estimates

$$\|Tf\|_{L^{q_\theta}(Y)} \leq B_\theta \|f\|_{L^{p_\theta}(X)} \tag{4.9}$$

for all $f \in L^{p_\theta}(X)$ and $0 \leq \theta \leq 1$, where the quantities $p_\theta, q_\theta, B_\theta$ are defined by the formulae

$$\frac{1}{p_\theta} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}$$
$$\frac{1}{q_\theta} = \frac{1-\theta}{q_0} + \frac{\theta}{q_1}$$
$$B_\theta = B_0^{1-\theta} B_1^\theta.$$

The Riesz-Thorin theorem is already quite useful (it gives, for instance, by far the quickest proof of the *Hausdorff-Young inequality* for the Fourier transform, to name just one application; see e.g.[**Ta2010**, (1.103)]), but it requires the *same* linear operator T to appear in (4.7), (4.8), and (4.9). Stein realised, though, that due to the complex-analytic nature of the proof of the Riesz-Thorin theorem, it was possible to allow *different* linear operators to appear in (4.7), (4.8), (4.9), so long as the dependence was analytic. A bit more precisely: if one had a family T_z of operators which depended in an analytic manner on a complex variable z in the strip $\{z \in \mathbf{C} : 0 \leq \mathrm{Re}(z) \leq 1\}$ (thus, for any test functions f, g, the inner product $\langle T_z f, g\rangle$ would be analytic in z) which obeyed some mild regularity assumptions (which are slightly technical and are omitted here), and one had the estimates

$$\|T_{0+it} f\|_{L^{q_0}(Y)} \leq C_t \|f\|_{L^{p_0}(X)}$$

and

$$\|T_{1+it} f\|_{L^{q_1}(Y)} \leq C_t \|f\|_{L^{p_1}(X)}$$

for all $t \in \mathbf{R}$ and some quantities C_t that grew at most exponentially in t (actually, any growth rate significantly slower than the double-exponential $e^{\exp(\pi|t|)}$ would suffice here), then one also has the interpolated estimates

$$\|T_\theta f\|_{L^{q_\theta}(Y)} \leq C' \|f\|_{L^{p_\theta}(X)}$$

for all $0 \leq \theta \leq 1$ and a constant C' depending only on C, p_0, p_1, q_0, q_1.

In [**Fe1995**], Fefferman notes that the proof of the Stein interpolation theorem can be obtained from that of the Riesz-Thorin theorem simply "by adding a single letter of the alphabet". Indeed, the way the Riesz-Thorin theorem is proven is to study an expression of the form

$$F(z) := \int_Y Tf_z(y) g_z(y)\, dy,$$

where f_z, g_z are functions depending on z in a suitably analytic manner, for instance taking $f_z = |f|^{\frac{1-z}{p_0} + \frac{z}{p_1}} \mathrm{sgn}(f)$ for some test function f, and

similarly for g. If f_z, g_z are chosen properly, F will depend analytically on z as well, and the two hypotheses (4.7), (4.8) give bounds on $F(0+it)$ and $F(1+it)$ for $t \in \mathbf{R}$ respectively. The *Lindelöf theorem* then gives bounds on intermediate values of F, such as $F(\theta)$; and the Riesz-Thorin theorem can then be deduced by a duality argument. (This is covered in many graduate real analysis texts; see e.g. [**Ta2010**, §1.11].)

The Stein interpolation theorem proceeds by instead studying the expression

$$F(z) := \int_Y T_z f_z(y) g_z(y)\ dy.$$

One can then repeat the proof of the Riesz-Thorin theorem more or less *verbatim* to obtain the Stein interpolation theorem.

The ability to vary the operator T makes the Stein interpolation theorem significantly more flexible than the Riesz-Thorin theorem. We illustrate this with the following sample result:

Proposition 4.3.1. *For any (test) function $f : \mathbf{R}^2 \to \mathbf{R}$, let $Tf : \mathbf{R}^2 \to \mathbf{R}$ be the average of f along an arc of a parabola:*

$$Tf(x_1, x_2) := \int_{\mathbf{R}} f(x_1 - t, x_2 - t^2)\eta(t)\ dt$$

where η is a bump function supported on (say) $[-1, 1]$. Then T is bounded from $L^{3/2}(\mathbf{R}^2)$ to $L^3(\mathbf{R}^2)$, thus

$$\|Tf\|_{L^3(\mathbf{R}^2)} \leq C\|f\|_{L^{3/2}(\mathbf{R}^2)}. \tag{4.10}$$

There is nothing too special here about the parabola; the same result in fact holds for convolution operators on any arc of a smooth curve with nonzero curvature (and there are many extensions to higher dimensions, to variable-coefficient operators, etc.). We will however restrict attention to the parabola for sake of exposition. One can view Tf as a convolution $Tf = f * \sigma$, where σ is a measure on the parabola arc $\{(t, t^2) : |t| \leq 1\}$. We will also be somewhat vague about what "test function" means in this exposition in order to gloss over some minor technical details.

By testing T (and its adjoint) on the indicator function of a small ball of some radius $\delta > 0$ (or of small rectangles such as $[-\delta, \delta] \times [0, \delta^2]$) one sees that the exponent $L^{3/2}$, L^3 here are best possible.

This proposition was first proven in [**Li1973**] using the Stein interpolation theorem. To illustrate the power of this theorem, it should be noted that for almost two decades this was the only known proof of this result; a proof based on multilinear interpolation (exploiting the fact that the exponent 3 in (4.10) is an integer) was obtained in [**Ob1992**], and a fully

combinatorial proof was only obtained in [**Ch2008**] (see also [**St2010**], [**DeFoMaWr2010**] for further extensions of the combinatorial argument).

To motivate the Stein interpolation argument, let us first try using the Riesz-Thorin interpolation theorem first. The exponent pair $L^{3/2} \to L^3$ is an interpolant between $L^2 \to L^2$ and $L^1 \to L^\infty$, so a first attempt to proceed here would be to establish the bounds

$$\|Tf\|_{L^2(\mathbf{R}^2)} \leq C\|f\|_{L^2(\mathbf{R}^2)} \tag{4.11}$$

and

$$\|Tf\|_{L^\infty(\mathbf{R}^2)} \leq C\|f\|_{L^1(\mathbf{R}^2)} \tag{4.12}$$

for all (test) functions f

The bound (4.11) is an easy consequence of *Minkowski's integral inequality*(or Young's inequality, noting that σ is a finite measure). On the other hand, because the measure σ is not absolutely continuous, let alone arising from an $L^\infty(\mathbf{R}^2)$ function, the estimate (4.12) is very false. For instance, if one applies Tf to the indicator function $1_{[-\delta,\delta]\times[-\delta,\delta]}$ for some small $\delta > 0$, then the L^1 norm of f is δ^2, but the L^∞ norm of Tf is comparable to δ, contradicting (4.12) as one sense δ to zero.

To get around this, one first notes that there is a lot of "room" in (4.11) due to the smoothing properties of the measure σ. Indeed, from Plancherel's theorem one has

$$\|f\|_{L^2(\mathbf{R}^2)} = \|\hat{f}\|_{L^2(\mathbf{R}^2)}$$

and

$$\|Tf\|_{L^2(\mathbf{R}^2)} = \|\hat{f}\hat{\sigma}\|_{L^2(\mathbf{R}^2)}$$

for all test functions f, where

$$\hat{f}(\xi) := \int_{\mathbf{R}^2} e^{-2\pi i x\cdot\xi} f(x)\ dx$$

is the Fourier transform of f, and

$$\hat{\sigma}(\xi_1,\xi_2) := \int_{\mathbf{R}} e^{-2\pi i(t\xi_1+t^2\xi_2)}\eta(t)\ dt.$$

It is clear that $\hat{\sigma}(\xi)$ is uniformly bounded in ξ, which already gives (4.11). But a standard application of the method of stationary phase reveals that one in fact has a decay estimate

$$|\hat{\sigma}(\xi)| \leq \frac{C}{|\xi|^{1/2}} \tag{4.13}$$

for some $C > 0$. This shows that Tf is not just in L^2, but is somewhat smoother as well; in particular, one has

$$\|D^{1/2}Tf\|_{L^2(\mathbf{R}^2)} \leq C\|f\|_{L^2(\mathbf{R}^2)}$$

for any (fractional) differential operator $D^{1/2}$ of order $1/2$. (Here we adopt the usual convention that the constant C is allowed to vary from line to line.)

Using the numerology of the Stein interpolation theorem, this suggests that if we can somehow obtain the counterbalancing estimate

$$\|D^{-1}Tf\|_{L^\infty(\mathbf{R}^2)} \leq C\|f\|_{L^1(\mathbf{R}^2)}$$

for some differential operator D^{-1} of order -1, then we should be able to interpolate and obtain the desired estimate (4.10). And indeed, we can take an antiderivative in the x_2 direction, giving the operator

$$\partial_{x_2}^{-1} Tf(x_1, x_2) := \int_{\mathbf{R}} \int_{-\infty}^{0} f(x_1 - t, x_2 - t^2 - s)\ \eta(t) dt ds;$$

and a simple change of variables does indeed verify that this operator is bounded from $L^1(\mathbf{R}^2)$ to $L^\infty(\mathbf{R}^2)$.

Unfortunately, the above argument is not rigorous, because we need an analytic family of operators T_z in order to invoke the Stein interpolation theorem, rather than just two operators T_0 and T_1. This turns out to require some slightly tricky complex analysis: after some trial and error, one finds that one can use the family T_z defined for $\mathrm{Re}(z) > 1/3$ by the formula

$$T_z f(x_1, x_2) = \frac{1}{\Gamma((3z-1)/2)} \int_{\mathbf{R}} \int_{-\infty}^{0} \frac{1}{s^{(3-3z)/2}} f(x_1 - t, x_2 - t^2 - s)\ \eta(t) dt ds$$

where Γ is the *Gamma function*, and extended to the rest of the complex plane by analytic continuation. The Gamma factor is a technical one, needed to compensate for the divergence of the weight $\frac{1}{s^{(3-3z)/2}}$ as z approaches $1/3$; it also makes the Fourier representation of T_z cleaner (indeed, $T_z f$ is morally $\partial_{x_2}^{(1-3z)/2} f * \sigma$). It is then easy to verify the estimates

$$\|T_{1+it} f\|_{L^\infty(\mathbf{R}^2)} \leq C_t \|f\|_{L^1(\mathbf{R}^2)} \tag{4.14}$$

for all $t \in \mathbf{R}$ (with C_t growing at a controlled rate), while from Fourier analysis one also can show that

$$\|T_{0+it} f\|_{L^2(\mathbf{R}^2)} \leq C_t \|f\|_{L^2(\mathbf{R}^2)} \tag{4.15}$$

for all $t \in \mathbf{R}$. Finally, one can verify that $T_{1/3} = T$, and (4.10) then follows from the Stein interpolation theorem.

It is instructive to compare this result with what can be obtained by real-variable methods. One can perform a smooth dyadic partition of unity

$$\delta(s) = \phi(s) + \sum_{j=1}^{\infty} 2^j \psi(2^j s)$$

for some bump function ϕ (of total mass 1) and bump function ψ (of total mass zero), which (formally, at least) leads to the decomposition

$$Tf = T_0 f + \sum_{j=1}^{\infty} T_j f$$

where $T_0 f$ is a harmless smoothing operator (which certainly maps $L^{3/2}(\mathbf{R}^2)$ to $L^3(\mathbf{R}^2)$) and

$$T_j f(x_1, x_2) := \int_{\mathbf{R}} \int_{\mathbf{R}} 2^j \psi(2^j s) f(x_1 - t, x_2 - t^2 - s) \eta(t)\, dt ds.$$

It is not difficult to show that

$$\|T_j f\|_{L^\infty(\mathbf{R}^2)} \leq C 2^j \|f\|_{L^1(\mathbf{R}^2)} \tag{4.16}$$

while a Fourier-analytic computation (using (4.13)) reveals that

$$\|T_j f\|_{L^2(\mathbf{R}^2)} \leq C 2^{-j/2} \|f\|_{L^2(\mathbf{R}^2)} \tag{4.17}$$

which interpolates (by, say, the Riesz-Thorin theorem, or the real-variable *Marcinkiewicz interpolation theorem*, see [**Ta2010**, Theorem 1.11.10]) to

$$\|T_j f\|_{L^3(\mathbf{R}^2)} \leq C \|f\|_{L^{3/2}(\mathbf{R}^2)}$$

which is close to (4.10). Unfortunately, we still have to sum in j, and this creates a "logarithmic divergence" that just barely fails[2] to recover (4.10).

The key difference is that the inputs (4.14), (4.15) used in the Stein interpolation theorem are more powerful than the inputs (4.16), (4.17) in the real-variable method. Indeed, (4.14) is roughly equivalent to the assertion that

$$\| \sum_{j=1}^{\infty} e^{2\pi i j t} 2^{-j} T_j f \|_{L^\infty(\mathbf{R}^2)} \leq C_t \|f\|_{L^1(\mathbf{R}^2)}$$

and (4.15) is similarly equivalent to the assertion that

$$\| \sum_{j=1}^{\infty} e^{2\pi i j t} 2^{j/2} T_j f \|_{L^2(\mathbf{R}^2)} \leq C_t \|f\|_{L^2(\mathbf{R}^2)}.$$

A Fourier averaging argument shows that these estimates imply (4.16) and (4.17), but not conversely. If one unpacks the proof of Lindelöf's theorem (which is ultimately powered by an integral representation, such as that provided by the *Cauchy integral formula*) and hence of the Stein interpolation theorem, one can interpret Stein interpolation in this case as using a clever integral representation of $\sum_{j=1}^{\infty} T_j f$ in terms of expressions such

[2] With a slightly more refined real interpolation argument, one can at least obtain a restricted weak-type estimate from $L^{3/2,1}(\mathbf{R}^2)$ to $L^{3,\infty}(\mathbf{R}^2)$ this way, but one can concoct abstract counterexamples to show that the estimates (4.16), (4.17) are insufficient to obtain an $L^{3/2} \to L^3$ bound on $\sum_{j=1}^{\infty} T_j$.

as $\sum_{j=1}^\infty e^{2\pi ijt} 2^{-j} T_j f_{1+it}$ and $\sum_{j=1}^\infty e^{2\pi ijt} 2^{j/2} T_j f_{0+it}$, where f_{1+it}, f_{0+it} are various nonlinear transforms of f. Technically, it would then be possible to rewrite the Stein interpolation argument as a real-variable one, without explicit mention of Lindelöf's theorem; but the proof would then look extremely contrived; the complex-analytic framework is much more natural (much as it is in analytic number theory, where the distribution of the primes is best handled by a complex-analytic study of the Riemann zeta function).

Remark 4.3.2. A useful strengthening of the Stein interpolation theorem is the *Fefferman-Stein interpolation theorem* [**FeSt1972**], in which the endpoint spaces L^1 and L^∞ are replaced by the *Hardy space* $\mathcal{H}^1$ and the space BMO of functions of *bounded mean oscillation* respectively. These spaces are more stable with respect to various harmonic analysis operators, such as singular integrals (and in particular, with respect to the Marcinkiewicz operators $|\nabla|^{it}$, which come up frequently when attempting to use the complex method), which makes the Fefferman-Stein theorem particularly useful for controlling expressions derived from these sorts of operators.

4.4. The Cotlar-Stein lemma

A basic problem in harmonic analysis (as well as in linear algebra, random matrix theory, and high-dimensional geometry) is to estimate the operator norm $\|T\|_{\mathrm{op}}$ of a linear map $T : H \to H'$ between two Hilbert spaces, which we will take to be complex for sake of discussion. Even the finite-dimensional case $T : \mathbf{C}^m \to \mathbf{C}^n$ is of interest, as this operator norm is the same as the largest *singular value* $\sigma_1(A)$ of the $n \times m$ matrix A associated to T.

In general, this operator norm is hard to compute precisely, except in special cases. One such special case is that of a *diagonal operator*, such as that associated to an $n \times n$ diagonal matrix $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. In this case, the operator norm is simply the supremum norm of the diagonal coefficients:

$$\|D\|_{\mathrm{op}} = \sup_{1 \leq i \leq n} |\lambda_i|. \tag{4.18}$$

A variant of (4.18) is *Schur's test*, which for simplicity we will phrase in the setting of finite-dimensional operators $T : \mathbf{C}^m \to \mathbf{C}^n$ given by a matrix $A = (a_{ij})_{1 \leq i \leq n; 1 \leq j \leq m}$ via the usual formula

$$T(x_j)_{j=1}^m := (\sum_{j=1}^m a_{ij} x_j)_{i=1}^n.$$

A simple version of this test is as follows: if all the absolute row sums and columns sums of A are bounded by some constant M, thus

$$\sum_{j=1}^{m} |a_{ij}| \leq M \tag{4.19}$$

for all $1 \leq i \leq n$ and

$$\sum_{i=1}^{n} |a_{ij}| \leq M \tag{4.20}$$

for all $1 \leq j \leq m$, then

$$\|T\|_{\mathrm{op}} = \|A\|_{\mathrm{op}} \leq M \tag{4.21}$$

(note that this generalises (the upper bound in) (4.18).) Indeed, to see (4.21), it suffices by duality and homogeneity to show that

$$|\sum_{i=1}^{n}(\sum_{j=1}^{m} a_{ij}x_j)y_i| \leq M$$

whenever $(x_j)_{j=1}^m$ and $(y_i)_{i=1}^n$ are sequences with $\sum_{j=1}^m |x_j|^2 = \sum_{i=1}^n |y_i|^2 = 1$; but this easily follows from the arithmetic mean-geometric mean inequality

$$|a_{ij}x_j)y_i| \leq \frac{1}{2}|a_{ij}||x_i|^2 + \frac{1}{2}|a_{ij}||y_j|^2$$

and (4.19), (4.20).

Schur's test (4.21) (and its many generalisations to weighted situations, or to Lebesgue or Lorentz spaces) is particularly useful for controlling operators in which the role of oscillation (as reflected in the *phase* of the coefficients a_{ij}, as opposed to just their magnitudes $|a_{ij}|$) is not decisive. However, it is of limited use in situations that involve a lot of cancellation. For this, a different test, known as the *Cotlar-Stein lemma* [**Co1955**], is much more flexible and powerful. It can be viewed in a sense as a non-commutative variant of Schur's test (4.21) (or of (4.18)), in which the scalar coefficients λ_i or a_{ij} are replaced by operators instead.

To illustrate the basic flavour of the result, let us return to the bound (4.18), and now consider instead a *block-diagonal* matrix

$$A = \begin{pmatrix} \Lambda_1 & 0 & \ldots & 0 \\ 0 & \Lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Lambda_n \end{pmatrix} \tag{4.22}$$

where each Λ_i is now a $m_i \times m_i$ matrix, and so A is an $m \times m$ matrix with $m := m_1 + \ldots + m_n$. Then we have

$$\|A\|_{\mathrm{op}} = \sup_{1 \leq i \leq n} \|\Lambda_i\|_{\mathrm{op}}. \tag{4.23}$$

Indeed, the lower bound is trivial (as can be seen by testing A on vectors which are supported on the i^{th} block of coordinates), while to establish the upper bound, one can make use of the orthogonal decomposition

$$\mathbf{C}^m \equiv \bigoplus_{i=1}^{m} \mathbf{C}^{m_i} \tag{4.24}$$

to decompose an arbitrary vector $x \in \mathbf{C}^m$ as

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

with $x_i \in \mathbf{C}^{m_i}$, in which case we have

$$Ax = \begin{pmatrix} \Lambda_1 x_1 \\ \Lambda_2 x_2 \\ \vdots \\ \Lambda_n x_n \end{pmatrix}$$

and the upper bound in (4.23) then follows from a simple computation.

The operator T associated to the matrix A in (4.22) can be viewed as a sum $T = \sum_{i=1}^n T_i$, where each T_i corresponds to the Λ_i block of A, in which case (4.23) can also be written as

$$\|T\|_{\mathrm{op}} = \sup_{1 \leq i \leq n} \|T_i\|_{\mathrm{op}}. \tag{4.25}$$

When n is large, this is a significant improvement over the triangle inequality, which merely gives

$$\|T\|_{\mathrm{op}} \leq \sum_{1 \leq i \leq n} \|T_i\|_{\mathrm{op}}.$$

The reason for this gain can ultimately be traced back to the "orthogonality" of the T_i; that they "occupy different columns" and "different rows" of the range and domain of T. This is obvious when viewed in the matrix formalism, but can also be described in the more abstract Hilbert space operator formalism via the identities[3]

$$T_i^* T_j = 0 \tag{4.26}$$

[3]The first identity asserts that the ranges of the T_i are orthogonal to each other, and the second asserts that the coranges of the T_i (the ranges of the adjoints T_i^*) are orthogonal to each other.

and

$$T_i T^* j = 0 \tag{4.27}$$

whenever $i \neq j$. By replacing (4.24) with a more abstract orthogonal decomposition into these ranges and coranges, one can in fact deduce (4.25) directly from (4.26) and (4.27).

The *Cotlar-Stein lemma* is an extension of this observation to the case where the T_i are merely *almost orthogonal* rather than *orthogonal*, in a manner somewhat analogous to how Schur's test (partially) extends (4.18) to the non-diagonal case. Specifically, we have

Lemma 4.4.1 (Cotlar-Stein lemma)**.** *Let $T_1, \ldots, T_n : H \to H'$ be a finite sequence of bounded linear operators from one Hilbert space H to another H', obeying the bounds*

$$\sum_{j=1}^{n} \|T_i T_j^*\|_{\mathrm{op}}^{1/2} \leq M \tag{4.28}$$

and

$$\sum_{j=1}^{n} \|T_i^* T_j\|_{\mathrm{op}}^{1/2} \leq M \tag{4.29}$$

for all $i = 1, \ldots, n$ and some $M > 0$ (compare with (4.19)*,* (4.20)*). Then one has*

$$\|\sum_{i=1}^{n} T_i\|_{\mathrm{op}} \leq M. \tag{4.30}$$

Note from the basic TT^* *identity*

$$\|T\|_{\mathrm{op}} = \|TT^*\|_{\mathrm{op}}^{1/2} = \|T^*T\|_{\mathrm{op}}^{1/2} \tag{4.31}$$

that the hypothesis (4.28) (or (4.29)) already gives the bound

$$\|T_i\|_{\mathrm{op}} \leq M \tag{4.32}$$

on each component T_i of T, which by the triangle inequality gives the inferior bound

$$\|\sum_{i=1}^{n} T_i\|_{\mathrm{op}} \leq nM;$$

the point of the Cotlar-Stein lemma is that the dependence on n in this bound is eliminated in (4.30), which in particular makes the bound suitable for extension to the limit $n \to \infty$ (see Remark 4.4.2 below).

The Cotlar-Stein lemma was first established by Cotlar [**Co1955**] in the special case of commuting self-adjoint operators, and then independently by Cotlar and Stein in full generality, with the proof appearing in [**KnSt1971**].

The Cotlar-Stein lemma is often useful in controlling operators such as *singular integral operators* or *pseudo-differential operators* T which "do not mix scales together too much", in that operators T map functions "that oscillate at a given scale 2^{-i}" to functions that still mostly oscillate at the same scale 2^{-i}. In that case, one can often split T into components T_i which essentically capture the scale 2^{-i} behaviour, and understanding L^2 boundedness properties of T then reduces to establishing the boundedness of the simpler operators T_i (and of establishing a sufficient decay in products such as $T_i^* T_j$ or $T_i T_j^*$ when i and j are separated from each other). In some cases, one can use Fourier-analytic tools such as Littlewood-Paley projections to generate the T_i, but the true power of the Cotlar-Stein lemma comes from situations in which the Fourier transform is not suitable, such as when one has a complicated domain (e.g. a manifold or a non-abelian Lie group), or very rough coefficients (which would then have badly behaved Fourier behaviour). One can then select the decomposition $T = \sum_i T_i$ in a fashion that is tailored to the particular operator T, and is not necessarily dictated by Fourier-analytic considerations.

Once one is in the almost orthogonal setting, as opposed to the genuinely orthogonal setting, the previous arguments based on orthogonal projection seem to fail completely. Instead, the proof of the Cotlar-Stein lemma proceeds via an elegant application of the *tensor power trick* (or perhaps more accurately, the *power method*), in which the operator norm of T is understood through the operator norm of a large power of T (or more precisely, of its self-adjoint square TT^* or T^*T). Indeed, from an iteration of (4.31) we see that for any natural number N, one has

$$\|T\|_{\mathrm{op}}^{2N} = \|(TT^*)^N\|_{\mathrm{op}}. \tag{4.33}$$

To estimate the right-hand side, we expand out the right-hand side and apply the triangle inequality to bound it by

$$\sum_{i_1,j_1,\ldots,i_N,j_N \in \{1,\ldots,n\}} \|T_{i_1} T_{j_1}^* T_{i_2} T_{j_2}^* \ldots T_{i_N} T_{j_N}^*\|_{\mathrm{op}}. \tag{4.34}$$

Recall that when we applied the triangle inequality directly to T, we lost a factor of n in the final estimate; it will turn out that we will lose a similar factor here, but this factor will eventually be attenuated into nothingness by the tensor power trick.

To bound (4.34), we use the basic inequality $\|ST\|_{\mathrm{op}} \leq \|S\|_{\mathrm{op}} \|T\|_{\mathrm{op}}$ in two different ways. If we group the product $T_{i_1} T_{j_1}^* T_{i_2} T_{j_2}^* \ldots T_{i_N} T_{j_N}^*$ in pairs, we can bound the summand of (4.34) by

$$\|T_{i_1} T_{j_1}^*\|_{\mathrm{op}} \ldots \|T_{i_N} T_{j_N}^*\|_{\mathrm{op}}.$$

On the other hand, we can group the product by pairs in another way, to obtain the bound of

$$\|T_{i_1}\|_{\mathrm{op}}\|T_{j_1}^* T_{i_2}\|_{\mathrm{op}} \dots \|T_{j_{N-1}}^* T_{i_N}\|_{\mathrm{op}}\|T_{j_N}^*\|_{\mathrm{op}}.$$

We bound $\|T_{i_1}\|_{\mathrm{op}}$ and $\|T_{j_N}^*\|_{\mathrm{op}}$ crudely by M using (4.32). Taking the geometric mean of the above bounds, we can thus bound (4.34) by

$$M \sum_{i_1,j_1,\dots,i_N,j_N \in \{1,\dots,n\}} \|T_{i_1} T_{j_1}^*\|_{\mathrm{op}}^{1/2}\|T_{j_1}^* T_{i_2}\|_{\mathrm{op}}^{1/2} \dots \|T_{j_{N-1}}^* T_{i_N}\|_{\mathrm{op}}^{1/2}\|T_{i_N} T_{j_N}^*\|_{\mathrm{op}}^{1/2}.$$

If we then sum this series first in j_N, then in i_N, then moving back all the way to i_1, using (4.28) and (4.29) alternately, we obtain a final bound of

$$nM^{2N}$$

for (4.33). Taking N^{th} roots, we obtain

$$\|T\|_{\mathrm{op}} \leq n^{1/2N} M.$$

Sending $N \to \infty$, we obtain the claim.

Remark 4.4.2. As observed in a number of places (see e.g. [**St1993**, p. 318] or [**Co2007**]), the Cotlar-Stein lemma can be extended to infinite sums $\sum_{i=1}^{\infty} T_i$ (with the obvious changes to the hypotheses (4.28), (4.29)). Indeed, one can show that for any $f \in H$, the sum $\sum_{i=1}^{\infty} T_i f$ is unconditionally convergent in H' (and furthermore has bounded 2-variation), and the resulting operator $\sum_{i=1}^{\infty} T_i$ is a bounded linear operator with an operator norm bound on M.

Remark 4.4.3. If we specialise to the case where all the T_i are equal, we see that the bound in the Cotlar-Stein lemma is sharp, at least in this case. Thus we see how the tensor power trick can convert an inefficient argument, such as that obtained using the triangle inequality or crude bounds such as (4.32), into an efficient one.

Remark 4.4.4. One can justify Schur's test by a similar method. Indeed, starting from the inequality

$$\|A\|_{\mathrm{op}}^{2N} \leq \mathrm{tr}((AA^*)^N)$$

(which follows easily from the singular value decomposition), we can bound $\|A\|_{\mathrm{op}}^{2N}$ by

$$\sum_{i_1,\dots,j_N \in \{1,\dots,n\}} a_{i_1,j_1}\overline{a_{j_1,i_2}} \dots a_{i_N,j_N}\overline{a_{j_N,i_1}}.$$

Estimating the other two terms in the summand by M, and then repeatedly summing the indices one at a time as before, we obtain

$$\|A\|_{\mathrm{op}}^{2N} \leq nM^{2N}$$

and the claim follows from the tensor power trick as before. On the other hand, in the converse direction, I do not know of any way to prove the Cotlar-Stein lemma that does not basically go through the tensor power argument.

4.5. Stein's spherical maximal inequality

If $f : \mathbf{R}^d \to \mathbf{C}$ is a locally integrable function, we define the *Hardy-Littlewood maximal function* $Mf : \mathbf{R}^d \to \mathbf{C}$ by the formula

$$Mf(x) := \sup_{r>0} \frac{1}{|B(x,r)|} \int_{B(x,r)} |f(y)|\ dy,$$

where $B(x,r)$ is the ball of radius r centred at x, and $|E|$ denotes the measure of a set E. The *Hardy-Littlewood maximal inequality* asserts that

$$|\{x \in \mathbf{R}^d : Mf(x) > \lambda\}| \leq \frac{C_d}{\lambda} \|f\|_{L^1(\mathbf{R}^d)} \tag{4.35}$$

for all $f \in L^1(\mathbf{R}^d)$, all $\lambda > 0$, and some constant $C_d > 0$ depending only on d. By a standard density argument, this implies in particular that we have the *Lebesgue differentiation theorem*

$$\lim_{r\to 0} \frac{1}{|B(x,r)|} \int_{B(x,r)} f(y)\ dy = f(x)$$

for all $f \in L^1(\mathbf{R}^d)$ and almost every $x \in \mathbf{R}^d$. See for instance [**Ta2011**, Theorem 1.6.11].

By combining the Hardy-Littlewood maximal inequality with the *Marcinkiewicz interpolation theorem* [**Ta2010**, 1.11.10] (and the trivial inequality $\|Mf\|_{L^\infty(\mathbf{R}^d)} \leq \|f\|_{L^\infty(\mathbf{R}^d)}$) we see that

$$\|Mf\|_{L^p(\mathbf{R}^d)} \leq C_{d,p} \|f\|_{L^p(\mathbf{R}^d)} \tag{4.36}$$

for all $p > 1$ and $f \in L^p(\mathbf{R}^d)$, and some constant $C_{d,p}$ depending on d and p.

The exact dependence of $C_{d,p}$ on d and p is still not completely understood. The standard Vitali-type covering argument used to establish (4.35) has an exponential dependence on dimension, giving a constant of the form $C_d = C^d$ for some absolute constant $C > 1$. Inserting this into the Marcinkiewicz theorem, one obtains a constant $C_{d,p}$ of the form $C_{d,p} = \frac{C^d}{p-1}$ for some $C > 1$ (and taking p bounded away from infinity, for simplicity). The dependence on p is about right, but the dependence on d should not be exponential.

In [**St1982, StSt1983**], Stein gave an elegant argument, based on the Calderón-Zygmund method of rotations, to eliminate the dependence of d:

Theorem 4.5.1. *One can take $C_{d,p} = C_p$ for each $p > 1$, where C_p depends only on p.*

The argument is based on an earlier bound [**St1976**] of Stein on the *spherical maximal function*

$$M_S f(x) := \sup_{r>0} A_r |f|(x)$$

where A_r are the spherical averaging operators

$$A_r f(x) := \int_{S^{d-1}} f(x + r\omega) d\sigma^{d-1}(\omega)$$

and $d\sigma^{d-1}$ is normalised surface measure on the sphere S^{d-1}. Because this is an uncountable supremum, and the averaging operators A_r do not have good continuity properties in r, it is not *a priori* obvious that $M_S f$ is even a measurable function for, say, locally integrable f; but we can avoid this technical issue, at least initially, by restricting attention to continuous functions f. The Stein maximal theorem for the spherical maximal function then asserts that if $d \geq 3$ and $p > \frac{d}{d-1}$, then we have

$$\|M_S f\|_{L^p(\mathbf{R}^d)} \leq C_{d,p} \|f\|_{L^p(\mathbf{R}^d)} \tag{4.37}$$

for all (continuous) $f \in L^p(\mathbf{R}^d)$. We will sketch a proof of this theorem[4] below the fold.

The condition $p > \frac{d}{d-1}$ can be seen to be necessary as follows. Take f to be any fixed bump function. A brief calculation then shows that $M_S f(x)$ decays like $|x|^{1-d}$ as $|x| \to \infty$, and hence $M_S f$ does not lie in $L^p(\mathbf{R}^d)$ unless $p > \frac{d}{d-1}$. By taking f to be a rescaled bump function supported on a small ball, one can show that the condition $p > \frac{d}{d-1}$ is necessary even if we replace $\mathbf{R}^d$ with a compact region (and similarly restrict the radius parameter r to be bounded). The condition $d \geq 3$ however is not quite necessary; the result is also true when $d = 2$, but this turned out to be a more difficult result, obtained first in [**Bo1985**], with a simplified proof (based on the local smoothing properties of the wave equation) later given in [**MoSeSo1992**].

The Hardy-Littlewood maximal operator Mf, which involves averaging over balls, is clearly related to the spherical maximal operator, which averages over spheres. Indeed, by using polar co-ordinates, one easily verifies the pointwise inequality

$$Mf(x) \leq M_S f(x)$$

for any (continuous) f, which intuitively reflects the fact that one can think of a ball as an average of spheres. Thus, we see that the spherical maximal

[4] Among other things, one can use this bound to show the pointwise convergence $\lim_{r\to 0} A_r f(x) = f(x)$ of the spherical averages for any $f \in L^p(\mathbf{R}^d)$ when $d \geq 3$ and $p > \frac{d}{d-1}$, although we will not focus on this application here.

inequality (4.37) implies[5] the Hardy-Littlewood maximal inequality (4.36) with the same constant $C_{p,d}$.

At first glance, this observation does not immediately establish Theorem 4.5.1 for two reasons. Firstly, Stein's spherical maximal theorem is restricted to the case when $d \geq 3$ and $p > \frac{d}{d-1}$; and secondly, the constant $C_{d,p}$ in that theorem still depends on dimension d. The first objection can be easily disposed of, for if $p > 1$, then the hypotheses $d \geq 3$ and $p > \frac{d}{d-1}$ will automatically be satisfied for d sufficiently large (depending on p); note that the case when d is bounded (with a bound depending on p) is already handled by the classical maximal inequality (4.36).

We still have to deal with the second objection, namely that constant $C_{d,p}$ in (4.37) depends on d. However, here we can use the method of rotations to show that the constants $C_{p,d}$ can be taken to be non-increasing (and hence bounded) in d. The idea is to view high-dimensional spheres as an average of rotated low-dimensional spheres. We illustrate this with a demonstration that $C_{d+1,p} \leq C_{d,p}$, in the sense that any bound of the form

$$\|M_S f\|_{L^p(\mathbf{R}^d)} \leq A\|f\|_{L^p(\mathbf{R}^d)} \tag{4.38}$$

for the d-dimensional spherical maximal function, implies the same bound

$$\|M_S f\|_{L^p(\mathbf{R}^{d+1})} \leq A\|f\|_{L^p(\mathbf{R}^{d+1})} \tag{4.39}$$

for the $d+1$-dimensional spherical maximal function, with exactly the same constant A. For any direction $\omega_0 \in S^d \subset \mathbf{R}^{d+1}$, consider the averaging operators

$$M_S^{\omega_0} f(x) := \sup_{r>0} A_r^{\omega_0} |f|(x)$$

for any continuous $f : \mathbf{R}^{d+1} \to \mathbf{C}$, where

$$A_r^{\omega_0} f(x) := \int_{S^{d-1}} f(x + rU_{\omega_0}\omega)\ d\sigma^{d-1}(\omega)$$

where U_{ω_0} is some orthogonal transformation mapping the sphere S^{d-1} to the sphere $S^{d-1,\omega_0} := \{\omega \in S^d : \omega \perp \omega_0\}$; the exact choice of orthogonal transformation U_{ω_0} is irrelevant due to the rotation-invariance of surface measure $d\sigma^{d-1}$ on the sphere S^{d-1}. A simple application of Fubini's theorem (after first rotating ω_0 to be, say, the standard unit vector e_d) using (4.38) then shows that

$$\|M_S^{\omega_0} f\|_{L^p(\mathbf{R}^{d+1})} \leq A\|f\|_{L^p(\mathbf{R}^{d+1})} \tag{4.40}$$

[5]This implication is initially only valid for continuous functions, but one can then extend the inequality (4.36) to the rest of $L^p(\mathbf{R}^d)$ by a standard limiting argument.

uniformly in ω_0. On the other hand, by viewing the d-dimensional sphere S^d as an average of the spheres S^{d-1,ω_0}, we have the identity

$$A_r f(x) = \int_{S^d} A_r^{\omega_0} f(x)\ d\sigma^d(\omega_0);$$

indeed, one can deduce this from the uniqueness of Haar measure by noting that both the left-hand side and right-hand side are invariant means of f on the sphere $\{y \in \mathbf{R}^{d+1} : |y-x| = r\}$. This implies that

$$M_S f(x) \leq \int_{S^d} M_S^{\omega_0} f(x)\ d\sigma^d(\omega_0)$$

and thus by Minkowski's inequality for integrals, we may deduce (4.39) from (4.40).

Remark 4.5.2. Unfortunately, the method of rotations does not work to show that the constant C_d for the weak $(1,1)$ inequality (4.35) is independent of dimension, as the weak L^1 quasinorm $\|\|_{L^{1,\infty}}$ is not a genuine norm and does not obey the Minkowski inequality for integrals. Indeed, the question of whether C_d in (4.35) can be taken to be independent of dimension remains open. The best known positive result is due to Stein and Strömberg [**StSt1983**], who showed that one can take $C_d = Cd$ for some absolute constant C, by comparing the Hardy-Littlewood maximal function with the heat kernel maximal function

$$\sup_{t>0} e^{t\Delta}|f|(x).$$

The abstract semigroup maximal inequality of Dunford and Schwartz (see e.g. [**Ta2009**, Theorem 2.9.1]) shows that the heat kernel maximal function is of weak-type $(1,1)$ with a constant of 1, and this can be used, together with a comparison argument, to give the Stein-Strömberg bound. In the converse direction, it was shown in [**Al2011**] that if one replaces the balls $B(x,r)$ with cubes, then the weak $(1,1)$ constant C_d must go to infinity as $d \to \infty$.

4.5.1. Proof of spherical maximal inequality. We now sketch the proof of Stein's spherical maximal inequality (4.37) for $d \geq 3$, $p > \frac{d}{d-1}$, and $f \in L^p(\mathbf{R}^d)$ continuous. To motivate the argument, let us first establish the simpler estimate

$$\|M_S^1 f\|_{L^p(\mathbf{R}^d)} \leq C_{d,p}\|f\|_{L^p(\mathbf{R}^d)}$$

where M_S^1 is the spherical maximal function restricted to unit scales:

$$M_S^1 f(x) := \sup_{1\leq r\leq 2} A_r|f|(x).$$

For the rest of these notes, we suppress the dependence of constants on d and p, using $X \lesssim Y$ as short-hand for $X \leq C_{p,d}Y$.

It will of course suffice to establish the estimate

$$\| \sup_{1\leq r\leq 2} |A_r f(x)| \|_{L^p(\mathbf{R}^d)} \lesssim \|f\|_{L^p(\mathbf{R}^d)} \tag{4.41}$$

for all continuous $f \in L^p(\mathbf{R}^d)$, as the original claim follows by replacing f with $|f|$. Also, since the bound is trivially true for $p = \infty$, and we crucially have $\frac{d}{d-1} < 2$ in three and higher dimensions, we can restrict attention to the regime $p < 2$.

We establish this bound using a Littlewood-Paley decomposition

$$f = \sum_N P_N f$$

where N ranges over dyadic numbers 2^k, $k \in \mathbf{Z}$, and P_N is a smooth Fourier projection to frequencies $|\xi| \sim N$; a bit more formally, we have

$$\widehat{P_N f}(\xi) = \psi(\frac{\xi}{N})\hat{f}(\xi)$$

where ψ is a bump function supported on the annulus $\{\xi \in \mathbf{R}^d : 1/2 \leq |\xi| \leq 2\}$ such that $\sum_N \psi(\frac{\xi}{N}) = 1$ for all non-zero ξ. Actually, for the purposes of proving (4.41), it is more convenient to use the decomposition

$$f = P_{\leq 1} f + \sum_{N>1} P_N f$$

where $P_{\leq 1} = \sum_{N\leq 1} P_N$ is the projection to frequencies $|\xi| \lesssim 1$. By the triangle inequality, it then suffices to show the bounds

$$\| \sup_{1\leq r\leq 2} |A_r P_{\leq 1} f(x)| \|_{L^p(\mathbf{R}^d)} \lesssim \|f\|_{L^p(\mathbf{R}^d)} \tag{4.42}$$

and

$$\| \sup_{1\leq r\leq 2} |A_r P_N f(x)| \|_{L^p(\mathbf{R}^d)} \lesssim N^{-\varepsilon} \|f\|_{L^p(\mathbf{R}^d)} \tag{4.43}$$

for all $N \geq 1$ and some $\varepsilon > 0$ depending only on p, d.

To prove the low-frequency bound (4.42), observe that $P_{\leq 1}$ is a convolution operator with a bump function, and from this and the radius restriction $1 \leq r \leq 2$ we see that $A_r P_{\leq 1}$ is a convolution operator with a function of uniformly bounded size and support. From this we obtain the pointwise bound

$$A_r P_{\leq 1} f(x) \lesssim Mf(x) \tag{4.44}$$

and the claim (4.42) follows from (4.36).

Now we turn to the more interesting high-frequency bound (4.43). Here, P_N is a convolution operator with an approximation to the identity at scale $\sim 1/N$, and so $A_r P_N$ is a convolution operator with a function of magnitude

$O(N)$ concentrated on an annulus of thickness $O(1/N)$ around the sphere of radius R. This can be used to give the pointwise bound

$$A_r P_N f(x) \lesssim N M f(x), \tag{4.45}$$

which by (4.36) gives the bound

$$\| \sup_{1 \leq r \leq 2} |A_r P_N f(x)| \|_{L^q(\mathbf{R}^d)} \lesssim_q N \|f\|_{L^q(\mathbf{R}^d)} \tag{4.46}$$

for any $q > 1$. This is not directly strong enough to prove (4.43), due to the "loss of one derivative" as manifested by the factor N. On the other hand, this bound (4.46) holds for all $q > 1$, and not just in the range $p > \frac{d}{d-1}$.

To counterbalance this loss of one derivative, we turn to L^2 estimates. A standard stationary phase computation (or Bessel function computation) shows that A_r is a Fourier multiplier whose symbol decays like $|\xi|^{-(d-1)/2}$. As such, Plancherel's theorem yields the L^2 bound

$$\|A_r P_N f\|_{L^2(\mathbf{R}^d)} \lesssim N^{-(d-1)/2} \|f\|_{L^2(\mathbf{R}^d)}$$

uniformly in $1 \leq r \leq 2$. But we still have to take the supremum over r. This is an uncountable supremum, so one cannot just apply a union bound argument. However, from the uncertainty principle, we expect $P_N f$ to be "blurred out" at spatial scale $1/N$, which suggests that the averages $A_r P_N f$ do not vary much when r is restricted to an interval of size $1/N$. Heuristically, this then suggests that

$$\sup_{1 \leq r \leq 2} |A_r P_N f| \sim \sup_{1 \leq r \leq 2 : r \in \frac{1}{N}\mathbf{Z}} |A_r P_N f|.$$

Estimating the discrete supremum on the right-hand side somewhat crudely by the square-function,

$$\sup_{1 \leq r \leq 2 : r \in \frac{1}{N}\mathbf{Z}} |A_r P_N f| \leq \Big(\sum_{1 \leq r \leq 2 : r \in \frac{1}{N}\mathbf{Z}} |A_r P_N f|^2 \Big)^{1/2},$$

and taking L^2 norms, one is then led to the heuristic prediction that

$$\| \sup_{1 \leq r \leq 2} |A_r P_N f| \|_{L^2(\mathbf{R}^d)} \lesssim N^{1/2} N^{-(d-1)/2} \|f\|_{L^2(\mathbf{R}^d)}. \tag{4.47}$$

One can make this heuristic precise using the one-dimensional Sobolev embedding inequality adapted to scale $1/N$, namely that

$$\sup_{1 \leq r \leq 2} |g(r)| \lesssim N^{1/2} \Big(\int_1^2 |g(r)|^2 \, dr \Big)^{1/2} + N^{-1/2} \Big(\int_1^2 |g'(r)|^2 \, dr \Big)^{1/2}.$$

A routine computation shows that

$$\| \frac{d}{dr} A_r P_N f \|_{L^2(\mathbf{R}^d)} \lesssim N \times N^{-(d-1)/2} \|f\|_{L^2(\mathbf{R}^d)}$$

(which formalises the heuristic that $A_r P_N f$ is roughly constant at r-scales $1/N$), and this soon leads to a rigorous proof of (4.47).

An interpolation between (4.46) and (4.47) (for q sufficiently close to 1) then gives (4.43) for some $\varepsilon > 0$ (here we crucially use that $p > \frac{d}{d-1}$ and $p < 2$).

Now we control the full maximal function $M_S f$. It suffices to show that

$$\| \sup_R \sup_{R \leq r \leq 2R} |A_r f(x)| \|_{L^p(\mathbf{R}^d)} \lesssim \|f\|_{L^p(\mathbf{R}^d)},$$

where R ranges over dyadic numbers.

For any fixed R, the natural spatial scale is R, and the natural frequency scale is thus $1/R$. We therefore split

$$f = P_{\leq 1/R} f + \sum_{N>1} P_{N/R} f,$$

and aim to establish the bounds

$$\| \sup_R \sup_{R \leq r \leq 2R} |A_r P_{\leq 1/R} f(x)| \|_{L^p(\mathbf{R}^d)} \lesssim \|f\|_{L^p(\mathbf{R}^d)} \tag{4.48}$$

and

$$\| \sup_R \sup_{R \leq r \leq 2R} |A_r P_{N/R} f(x)| \|_{L^p(\mathbf{R}^d)} \lesssim N^{-\varepsilon} \|f\|_{L^p(\mathbf{R}^d)} \tag{4.49}$$

for each $N > 1$ and some $\varepsilon > 0$ depending only on d and p, similarly to before.

A rescaled version of the derivation of (4.44) gives

$$A_r P_{\leq 1/R} f(x) \lesssim M f(x)$$

for all $R \leq r \leq 2R$, which already lets us deduce (4.48). As for (4.49), a rescaling of (4.45) gives

$$A_r P_{N/R} f(x) \lesssim N M f(x),$$

for all $R \leq r \leq 2R$, and thus

$$\| \sup_R \sup_{R \leq r \leq 2R} |A_r P_{N/R} f(x)| \|_{L^q(\mathbf{R}^d)} \lesssim N \|f\|_{L^q(\mathbf{R}^d)} \tag{4.50}$$

for all $q > 1$. Meanwhile, at the L^2 level, we have

$$\|A_r P_{N/R} f\|_{L^2(\mathbf{R}^d)} \lesssim N^{-(d-1)/2} \|f\|_{L^2(\mathbf{R}^d)}$$

and

$$\| \frac{d}{dr} A_r P_{N/R} f\|_{L^2(\mathbf{R}^d)} \lesssim \frac{N}{R} N^{-(d-1)/2} \|f\|_{L^2(\mathbf{R}^d)}$$

and so

$$\|(\frac{1}{R}\int_R^{2R} |A_r P_{N/R} f|^2\ dr)^{1/2} + (\frac{R}{N^2}\int_R^{2R} |\frac{d}{dr} A_r P_{N/R} f|^2\ dr)^{1/2}\|_{L^2(\mathbf{R}^d)}$$
$$\lesssim N^{1/2} N^{-(d-1)/2} \|f\|_{L^2(\mathbf{R}^d)}$$

which implies by rescaled Sobolev embedding that

$$\| \sup_{R\leq r\leq 2R} |A_r P_{N/R} f|\|_{L^2(\mathbf{R}^d)} \lesssim N^{1/2} N^{-(d-1)/2} \|f\|_{L^2(\mathbf{R}^d)}.$$

In fact, by writing $P_{N/R} f = P_{N/R} \tilde{P}_{N/R} f$, where $\tilde{P}_{N/R}$ is a slight widening of $P_{N/R}$, we have

$$\| \sup_{R\leq r\leq 2R} |A_r P_{N/R} f|\|_{L^2(\mathbf{R}^d)} \lesssim N^{1/2} N^{-(d-1)/2} \|\tilde{P}_{N/R} f\|_{L^2(\mathbf{R}^d)};$$

square summing this (and bounding a supremum by a square function) and using Plancherel we obtain

$$\| \sup_R \sup_{R\leq r\leq 2R} |A_r P_{N/R} f|\|_{L^2(\mathbf{R}^d)} \lesssim N^{1/2} N^{-(d-1)/2} \|f\|_{L^2(\mathbf{R}^d)}.$$

Interpolating this against (4.50) as before we obtain (4.49) as required.

4.6. Stein's maximal principle

Suppose one has a measure space $X = (X, \mathcal{B}, \mu)$ and a sequence of operators $T_n : L^p(X) \to L^p(X)$ that are bounded on some $L^p(X)$ space, with $1 \leq p < \infty$. Suppose that on some dense subclass of functions f in $L^p(X)$ (e.g. continuous compactly supported functions, if the space X is reasonable), one already knows that $T_n f$ converges pointwise almost everywhere to some limit Tf, for another bounded operator $T : L^p(X) \to L^p(X)$ (e.g. T could be the identity operator). What additional ingredient does one need to pass to the limit and conclude that $T_n f$ converges almost everywhere to Tf for *all* f in $L^p(X)$ (and not just for f in a dense subclass)?

One standard way to proceed here is to study the *maximal operator*

$$T_* f(x) := \sup_n |T_n f(x)|$$

and aim to establish a *weak-type maximal inequality*

$$\|T_* f\|_{L^{p,\infty}(X)} \leq C \|f\|_{L^p(X)} \tag{4.51}$$

for all $f \in L^p(X)$ (or all f in the dense subclass), and some constant C, where $L^{p,\infty}$ is the weak L^p norm

$$\|f\|_{L^{p,\infty}(X)} := \sup_{t>0} t\mu(\{x \in X : |f(x)| \geq t\})^{1/p}.$$

A standard approximation argument using (4.51) then shows that $T_n f$ will now indeed converge to Tf pointwise almost everywhere for all f in $L^p(X)$,

and not just in the dense subclass. See for instance [**Ta2011**, §1.6], in which this method is used to deduce the *Lebesgue differentiation theorem* from the *Hardy-Littlewood maximal inequality*.

This is by now a very standard approach to establishing pointwise almost everywhere convergence theorems, but it is natural to ask whether it is strictly necessary. In particular, is it possible to have a pointwise convergence result $T_n f \mapsto Tf$ without being able to obtain a weak-type maximal inequality of the form (4.51)?

In the case of *norm* convergence (in which one asks for $T_n f$ to converge to Tf in the L^p norm, rather than in the pointwise almost everywhere sense), the answer is no, thanks to the *uniform boundedness principle*, which among other things shows that norm convergence is only possible if one has the uniform bound

$$\sup_n \|T_n f\|_{L^p(X)} \leq C\|f\|_{L^p(X)} \tag{4.52}$$

for some $C > 0$ and all $f \in L^p(X)$; and conversely, if one has the uniform bound, and one has already established norm convergence of $T_n f$ to Tf on a dense subclass of $L^p(X)$, (4.52) will extend that norm convergence to all of $L^p(X)$.

Returning to pointwise almost everywhere convergence, the answer in general is "yes". Consider for instance the rank one operators

$$T_n f(x) := 1_{[n,n+1]} \int_0^1 f(y)\ dy$$

from $L^1(\mathbf{R})$ to $L^1(\mathbf{R})$. It is clear that $T_n f$ converges pointwise almost everywhere to zero as $n \to \infty$ for any $f \in L^1(\mathbf{R})$, and the operators T_n are uniformly bounded on $L^1(\mathbf{R})$, but the maximal function T_* does not obey (4.51). One can modify this example in a number of ways to defeat almost any reasonable conjecture that something like (4.51) should be necessary for pointwise almost everywhere convergence.

In spite of this, a remarkable observation of Stein [**St1961**], now known as *Stein's maximal principle*, asserts that the maximal inequality *is* necessary to prove pointwise almost everywhere convergence, if one is working on a compact group and the operators T_n are translation invariant, and if the exponent p is at most 2:

Theorem 4.6.1 (Stein maximal principle)**.** *Let G be a compact group, let X be a* homogeneous space[6] *of G with a finite Haar measure μ, let $1 \leq p \leq 2$, and let $T_n : L^p(X) \to L^p(X)$ be a sequence of bounded linear operators commuting with translations, such that $T_n f$ converges pointwise almost everywhere for each $f \in L^p(X)$. Then* (4.51) *holds.*

[6]By this, we mean that G has a transitive action on X which preserves μ.

This is not quite the most general version of the principle; some additional variants and generalisations are given in [**St1961**]. For instance, one can replace the discrete sequence T_n of operators with a continuous sequence T_t without much difficulty. As a typical application of this principle, we see that Carleson's celebrated theorem [**Ca1966**] that the partial Fourier series $\sum_{n=-N}^{N} \hat{f}(n) e^{2\pi i n x}$ of an $L^2(\mathbf{R}/\mathbf{Z})$ function $f : \mathbf{R}/\mathbf{Z} \to \mathbf{C}$ converge almost everywhere is in fact equivalent to the estimate

$$\| \sup_{N>0} | \sum_{n=-N}^{N} \hat{f}(n) e^{2\pi i n \cdot} | \|_{L^{2,\infty}(\mathbf{R}/\mathbf{Z})} \leq C \|f\|_{L^2(\mathbf{R}/\mathbf{Z})}. \tag{4.53}$$

And unsurprisingly, most of the proofs of this (difficult) theorem have proceeded by first establishing (4.53), and Stein's maximal principle strongly suggests that this is the optimal way to try to prove this theorem.

On the other hand, the theorem does fail for $p > 2$, and almost everywhere convergence results in L^p for $p > 2$ can be proven by other methods than weak (p,p) estimates. For instance, the convergence of Bochner-Riesz multipliers in $L^p(\mathbf{R}^n)$ for any n (and for p in the range predicted by the Bochner-Riesz conjecture) was verified for $p > 2$ in [**CaRuVe1988**], despite the fact that the weak (p,p) of even a *single* Bochner-Riesz multiplier, let alone the maximal function, has still not been completely verified in this range. (The argument in [**CaRuVe1988**] uses weighted L^2 estimates for the maximal Bochner-Riesz operator, rather than L^p type estimates.) For $p \leq 2$, though, Stein's principle (after localising to a torus) does apply, though, and pointwise almost everywhere convergence of Bochner-Riesz means is equivalent to the weak (p,p) estimate (4.51).

Stein's principle is restricted to compact groups (such as the torus $(\mathbf{R}/\mathbf{Z})^n$ or the rotation group $SO(n)$) and their homogeneous spaces (such as the torus $(\mathbf{R}/\mathbf{Z})^n$ again, or the sphere S^{n-1}). As stated, the principle fails in the noncompact setting; for instance, in $\mathbf{R}$, the convolution operators $T_n f := f * 1_{[n,n+1]}$ are such that $T_n f$ converges pointwise almost everywhere to zero for every $f \in L^1(\mathbf{R}^n)$, but the maximal function is not of weak-type $(1,1)$. However, in many applications on non-compact domains, the T_n are "localised" enough that one can transfer from a non-compact setting to a compact setting and then apply Stein's principle. For instance, Carleson's theorem on the real line $\mathbf{R}$ is equivalent to Carleson's theorem on the circle $\mathbf{R}/\mathbf{Z}$ (due to the localisation of the Dirichlet kernels), which as discussed before is equivalent to the estimate (4.53) on the circle, which by a scaling argument is equivalent to the analogous estimate on the real line $\mathbf{R}$.

Stein's argument from [**St1961**] can be viewed nowadays as an application of the *probabilistic method*; starting with a sequence of increasingly bad counterexamples to the maximal inequality (4.51), one randomly combines

them together to create a single "infinitely bad" counterexample. To make this idea work, Stein employs two basic ideas:

(1) The *random rotations (or random translations) trick*. Given a subset E of X of small but positive measure, one can randomly select about $|G|/|E|$ translates $g_i E$ of E that cover most of X.

(2) The *random sums trick* Given a collection $f_1, \ldots, f_n : X \to \mathbf{C}$ of signed functions that may possibly cancel each other in a deterministic sum $\sum_{i=1}^n f_i$, one can perform a random sum $\sum_{i=1}^n \pm f_i$ instead to obtain a random function whose magnitude will usually be comparable to the square function $(\sum_{i=1}^n |f_i|^2)^{1/2}$; this can be made rigorous by *concentration of measure* results, such as *Khintchine's inequality*.

These ideas have since been used repeatedly in harmonic analysis. For instance, the random rotations trick was used in [**ElObTa2010**] to obtain Kakeya-type estimates in finite fields. The random sums trick is by now a standard tool to build various counterexamples to estimates (or to convergence results) in harmonic analysis, for instance being used in [**Fe1971**] to disprove the boundedness of the ball multiplier on $L^p(\mathbf{R}^n)$ for $p \neq 2$, $n \geq 2$. Another use of the random sum trick is to show that Theorem 4.6.1 fails once $p > 2$; see Stein's original paper for details.

Another use of the random rotations trick, closely related to Theorem 4.6.1, is the *Nikishin-Stein factorisation theorem*. Here is Stein's formulation of this theorem:

Theorem 4.6.2 (Stein factorisation theorem)**.** *Let G be a compact group, let X be a homogeneous space of G with a finite Haar measure μ, let $1 \leq p \leq 2$ and $q > 0$, and let $T : L^p(X) \to L^q(X)$ be a bounded linear operator commuting with translations and obeying the estimate*

$$\|Tf\|_{L^q(X)} \leq A\|f\|_{L^p(X)}$$

for all $f \in L^p(X)$ and some $A > 0$. Then T also maps $L^p(X)$ to $L^{p,\infty}(X)$, with

$$\|Tf\|_{L^{p,\infty}(X)} \leq C_{p,q} A\|f\|_{L^p(X)}$$

for all $f \in L^p(X)$, with $C_{p,q}$ depending only on p, q.

This result is trivial with $q \geq p$, but becomes useful when $q < p$. In this regime, the translation invariance allows one to freely "upgrade" a strong-type (p, q) result to a weak-type (p, p) result. In other words, bounded linear operators from $L^p(X)$ to $L^q(X)$ automatically factor through the inclusion $L^{p,\infty}(X) \subset L^q(X)$, which helps explain the name "factorisation theorem". Factorisation theory has been developed further in [**Ma1974**], [**Pi1986**].

Stein's factorisation theorem (or more precisely, a variant of it) is useful in the theory of Kakeya and restriction theorems in Euclidean space, as first observed in [**Bo1991**].

In [**Ni1970**], Nikishin obtained the following generalisation of Stein's factorisation theorem in which the translation-invariance hypothesis can be dropped, at the cost of excluding a set of small measure:

Theorem 4.6.3 (Nikishin-Stein factorisation theorem)**.** *Let X be a finite measure space, let $1 \leq p \leq 2$ and $q > 0$, and let $T : L^p(X) \to L^q(X)$ be a bounded linear operator commuting with translations and obeying the estimate*

$$\|Tf\|_{L^q(X)} \leq A\|f\|_{L^p(X)}$$

for all $f \in L^p(X)$ and some $A > 0$. Then for any $\varepsilon > 0$, there exists a subset E of X of measure at most ε such that

$$\|Tf\|_{L^{p,\infty}(X\backslash E)} \leq C_{p,q,\varepsilon} A\|f\|_{L^p(X)} \tag{4.54}$$

for all $f \in L^p(X)$, with $C_{p,q,\varepsilon}$ depending only on p, q, ε.

One can recover Theorem 4.6.2 from Theorem 4.6.3 by an averaging argument to eliminate the exceptional set; we omit the details.

4.6.1. Sketch of proofs. We now sketch how Stein's maximal principle is proven. We may normalise $\mu(X) = 1$. Suppose the maximal inequality (4.51) fails for any C. Then, for any $A \geq 1$, we can find a non-zero function $f \in L^p(X)$ such that

$$\|T_* f\|_{L^{p,\infty}(X)} \geq A\|f\|_{L^p(X)}.$$

By homogeneity, we can arrange matters so that

$$\mu(E) \geq A^p \|f\|^p_{L^p(X)},$$

where $E := \{x \in X : |T_* f(x)| \geq 1\}$.

At present, E could be a much smaller set than X: $\mu(E) \ll 1$. But we can amplify E by using the random rotations trick. Let m be a natural number comparable to $1/\mu(E)$, and let $g_1, \ldots, g_m$ be elements of G, chosen uniformly at random. Each element x of X has a probability $1 - (1 - \mu(E))^m \sim 1$ of lying in at least one of the translates $g_1 E, \ldots, g_m E$ of E. From this and the first moment method, we see that with probability ~ 1, the set $g_1 E \cup \ldots \cup g_m E$ has measure ~ 1.

Now form the function $F := \sum_{j=1}^m \varepsilon_j \tau_{g_j} f$, where $\tau_{g_j} f(x) := f(g_j^{-1} x)$ is the left-translation of f by g_j, and the $\varepsilon_j = \pm 1$ are randomly chosen signs. On the one hand, an application of moment methods (such as the *Paley-Zygmund inequality*), one can show that each element x of $g_1 E \cup \ldots \cup g_m E$ will be such that $|T_* F(x)| \gtrsim 1$ with probability ~ 1. On the other hand,

an application of Khintchine's inequality shows that with high probability F will have an $L^p(X)$ norm bounded by

$$\lesssim \|(\sum_{j=1}^{m} |\tau_{g_j} f|^2)^{1/2}\|_{L^p(X)}.$$

Now we crucially use the hypothesis $p \leq 2$ to replace the ℓ^2-summation here by an ℓ^p summation. Interchanging the ℓ^p and L^p norms, we then conclude that with high probability we have

$$\|F\|_{L^p(X)} \lesssim m^{1/p} \|f\|_{L^p(X)} \lesssim 1/A.$$

To summarise, using the probabilistic method, we have constructed (for arbitrarily large A) a function $F = F_A$ whose L^p norm is only $O(1/A)$ in size, but such that $|T_* F(x)| \gtrsim 1$ on a subset of X of measure ~ 1. By sending A rapidly to infinity and taking a suitable combination of these functions F, one can then create a function G in L^p such that $T_* G$ is infinite on a set of positive measure, which contradicts the hypothesis of pointwise almost everywhere convergence.

Stein's factorisation theorem is proven in a similar fashion. For Nikishin's factorisation theorem, the group translation operations τ_{g_j} are no longer available. However, one can substitute for this by using the failure of the hypothesis (4.54), which among other things tells us that if one has a number of small sets $E_1, \ldots, E_i$ in X whose total measure is at most ε, then we can find another function f_{i+1} of small L^p norm for which $T f_{i+1}$ is large on a set E_{i+1} outside of $E_1 \cup \ldots \cup E_i$. Iterating this observation and choosing all parameters carefully, one can eventually establish the result.

Remark 4.6.4. A systematic discussion of these and other maximal principles is given in [**de1981**].

Chapter 5

Nonstandard analysis

5.1. Polynomial bounds via nonstandard analysis

Nonstandard analysis is useful in allowing one to import tools from infinitary (or qualitative) mathematics in order to establish results in finitary (or quantitative) mathematics. One drawback, though, to using nonstandard analysis methods is that the bounds one obtains by such methods are usually *ineffective*: in particular, the conclusions of a nonstandard analysis argument may involve an unspecified constant C that is known to be finite but for which no explicit bound is obviously[1] available.

Because of this fact, it would seem that quantitative bounds, such as polynomial type bounds $X \leq CY^C$ that show that one quantity X is controlled in a polynomial fashion by another quantity Y, are not easily obtainable through the ineffective methods of nonstandard analysis. Actually, this is not the case; as I will demonstrate by an example below, nonstandard analysis can certainly yield polynomial type bounds. The catch is that the exponent C in such bounds will be ineffective; but nevertheless such bounds are still good enough for many applications.

Let us now illustrate this by reproving a lemma of Chang [**Ch2003**] (Lemma 2.14, to be precise), which was recently pointed out to me by Van Vu. Chang's paper is focused primarily on the sum-product problem, but she uses a quantitative lemma from algebraic geometry which is of independent

[1] In many cases, a bound can eventually be worked out by performing *proof mining* on the argument, and in particular by carefully unpacking the *proofs* of all the various results from infinitary mathematics that were used in the argument, as opposed to simply using them as "black boxes", but this is a time-consuming task and the bounds that one eventually obtains tend to be quite poor (e.g. tower exponential or Ackermann type bounds are not uncommon).

interest. To motivate the lemma, let us first establish a qualitative version (a variant of the *Lefschetz principle*):

Lemma 5.1.1 (Qualitative solvability)**.** *Let $P_1, \ldots, P_r : \mathbf{C}^d \to \mathbf{C}$ be a finite number of polynomials in several variables with rational coefficients. If there is a complex solution $z = (z_1, \ldots, z_d) \in \mathbf{C}^d$ to the simultaneous system of equations*

$$P_1(z) = \ldots = P_r(z) = 0,$$

then there also exists a solution $z \in \overline{\mathbf{Q}}^d$ whose coefficients are algebraic numbers (i.e. they lie in the algebraic closure $\mathbf{Q}$ of the rationals).

Proof. Suppose there was no solution to $P_1(z) = \ldots = P_r(z) = 0$ over $\overline{\mathbf{Q}}$. Applying *Hilbert's nullstellensatz* (which is available as $\overline{\mathbf{Q}}$ is algebraically closed), we conclude the existence of some polynomials $Q_1, \ldots, Q_r$ (with coefficients in $\overline{\mathbf{Q}}$) such that

$$P_1 Q_1 + \ldots + P_r Q_r = 1$$

as polynomials. In particular, we have

$$P_1(z)Q_1(z) + \ldots + P_r(z)Q_r(z) = 1$$

for all $z \in \mathbf{C}^d$. This shows that there is no solution to $P_1(z) = \ldots = P_r(z) = 0$ over $\mathbf{C}$, as required. □

Remark 5.1.2. Observe that in the above argument, one could replace $\mathbf{Q}$ and $\mathbf{C}$ by any other pair of fields, with the latter containing the algebraic closure of the former, and still obtain the same result.

The above lemma asserts that if a system of rational equations is solvable at all, then it is solvable with some algebraic solution. But it gives no bound on the complexity of that solution in terms of the complexity of the original equation. Chang's lemma provides such a bound. If $H \geq 1$ is an integer, let us say that an algebraic number has *height* at most H if its minimal polynomial (after clearing denominators) consists of integers of magnitude at most H.

Lemma 5.1.3 (Quantitative solvability)**.** *Let $P_1, \ldots, P_r : \mathbf{C}^d \to \mathbf{C}$ be a finite number of polynomials of degree at most D with rational coefficients, each of height at most H. If there is a complex solution $z = (z_1, \ldots, z_d) \in \mathbf{C}^d$ to the simultaneous system of equations*

$$P_1(z) = \ldots = P_r(z) = 0,$$

then there also exists a solution $z \in \overline{\mathbf{Q}}^d$ whose coefficients are algebraic numbers of degree at most C and height at most CH^C, where $C = C_{D,d,r}$ depends only on D, d and r.

Chang proves this lemma by essentially establishing a quantitative version of the nullstellensatz, via elementary elimination theory (somewhat similar, actually, to the approach taken in I took to the nullstellensatz in [**Ta2008**, §1.15]. She also notes that one could also establish the result through the machinery of *Gröbner bases*. In each of these arguments, it was not possible to use Lemma 5.1.1 (or the closely related nullstellensatz) as a black box; one actually had to unpack one of the proofs of that lemma or nullstellensatz to get the polynomial bound. However, using nonstandard analysis, it is possible to get such polynomial bounds (albeit with an ineffective value of the constant C) directly from Lemma 5.1.1 (or more precisely, the generalisation in Remark 5.1.2) *without* having to inspect the proof, and instead simply using it as a black box, thus providing a "soft" proof of Lemma 5.1.3 that is an alternative to the "hard" proofs mentioned above.

The nonstandard proof is essentially due to Schmidt-Göttsch [**Sc1989**], and proceeds as follows. Informally, the idea is that Lemma 5.1.3 should follow from Lemma 5.1.1 after replacing the field of rationals $\mathbf{Q}$ with "the field of rationals of polynomially bounded height". Unfortunately, the latter object does not really make sense as a field in standard analysis; nevertheless, it is a perfectly sensible object in nonstandard analysis, and this allows the above informal argument to be made rigorous.

We turn to the details. As is common whenever one uses nonstandard analysis to prove finitary results, we use a "compactness and contradiction" argument (or more precisely, an "ultralimit and contradiction" argument). Suppose for contradiction that Lemma 5.1.3 failed. Carefully negating the quantifiers (and using the axiom of choice), we conclude that there exists D, d, r such that for each natural number n, there is a positive integer $H^{(n)}$ and a family $P_1^{(n)}, \ldots, P_r^{(n)} : \mathbf{C}^d \to \mathbf{C}$ of polynomials of degree at most D and rational coefficients of height at most $H^{(n)}$, such that there exist at least one complex solution $z^{(n)} \in \mathbf{C}^d$ to

$$P_1^{(n)}(z^{(n)}) = \ldots = P_r(z^{(n)}) = 0, \tag{5.1}$$

but such that there does not exist any such solution whose coefficients are algebraic numbers of degree at most n and height at most $n(H^{(n)})^n$.

Now we take ultralimits (see e.g. [**Ta2011b**, §2.1] of a quick review of ultralimit analysis, which we will assume knowledge of in the argument that follows). Let $p \in \beta\mathbf{N}\backslash\mathbf{N}$ be a non-principal *ultrafilter*. For each $i = 1, \ldots, r$, the ultralimit

$$P_i := \lim_{n \to p} P_i^{(n)}$$

of the (standard) polynomials $P_i^{(n)}$ is a nonstandard polynomial $P_i : {}^*\mathbf{C}^d \to {}^*\mathbf{C}$ of degree at most D, whose coefficients now lie in the nonstandard rationals ${}^*\mathbf{Q}$. Actually, due to the height restriction, we can say more. Let

$H := \lim_{n\to p} H^{(n)} \in {}^*\mathbf{N}$ be the ultralimit of the $H^{(n)}$, this is a nonstandard natural number (which will almost certainly be unbounded, but we will not need to use this). Let us say that a nonstandard integer a is *of polynomial size* if we have $|a| \leq CH^C$ for some standard natural number C, and say that a nonstandard rational number a/b is *of polynomial height* if a, b are of polynomial size. Let $\mathbf{Q}_{\mathrm{poly}(H)}$ be the collection of all nonstandard rationals of polynomial height. (In the language of nonstandard analysis, $\mathbf{Q}_{\mathrm{poly}(H)}$ is an *external* set rather than an internal one, because it is not itself an ultraproduct of standard sets; but this will not be relevant for the argument that follows.) It is easy to see that $\mathbf{Q}_{\mathrm{poly}(H)}$ is a field, basically because the sum or product of two integers of polynomial size, remains of polynomial size. By construction, it is clear that the coefficients of P_i are nonstandard rationals of polynomial height, and thus $P_1, \ldots, P_r$ are defined over $\mathbf{Q}_{\mathrm{poly}(H)}$.

Meanwhile, if we let $z := \lim_{n\to p} z^{(n)} \in {}^*\mathbf{C}^d$ be the ultralimit of the solutions $z^{(n)}$ in (5.1), we have

$$P_1(z) = \ldots = P_r(z) = 0,$$

thus $P_1, \ldots, P_r$ are solvable in ${}^*\mathbf{C}$. Applying Lemma 5.1.1 (or more precisely, the generalisation in Remark 5.1.2), we see that $P_1, \ldots, P_r$ are also solvable in $\overline{\mathbf{Q}_{\mathrm{poly}(H)}}$. (Note that as $\mathbf{C}$ is algebraically closed, ${}^*\mathbf{C}$ is also (by Los's theorem), and so ${}^*\mathbf{C}$ contains $\overline{\mathbf{Q}_{\mathrm{poly}(H)}}$.) Thus, there exists $w \in \overline{\mathbf{Q}_{\mathrm{poly}(H)}}^d$ with

$$P_1(w) = \ldots = P_r(w) = 0.$$

As $\overline{\mathbf{Q}_{\mathrm{poly}(H)}}^d$ lies in ${}^*\mathbf{C}^d$, we can write w as an ultralimit $w = \lim_{n\to p} w^{(n)}$ of standard complex vectors $w^{(n)} \in \mathbf{C}^d$. By construction, the coefficients of w each obey a non-trivial polynomial equation of degree at most C and whose coefficients are nonstandard integers of magnitude at most CH^C, for some standard natural number C. Undoing the ultralimit, we conclude that for n sufficiently close to p, the coefficients of $w^{(n)}$ obey a non-trivial polynomial equation of degree at most C whose coefficients are *standard* integers of magnitude at most $C(H^{(n)})^C$. In particular, these coefficients have height at most $C(H^{(n)})^C$. Also, we have

$$P_1^{(n)}(w^{(n)}) = \ldots = P_r^{(n)}(w^{(n)}) = 0.$$

But for n larger than C, this contradicts the construction of the $P_i^{(n)}$, and the claim follows. (Note that as p is non-principal, any neighbourhood of p in $\mathbf{N}$ will contain arbitrarily large natural numbers.)

Remark 5.1.4. The same argument actually gives a slightly stronger version of Lemma 5.1.3, namely that the integer coefficients used to define the algebraic solution z can be taken to be polynomials in the coefficients of $P_1, \ldots, P_r$, with degree and coefficients bounded by $C_{D,d,r}$.

5.2. Loeb measure and the triangle removal lemma

Formally, a *measure space* is a triple $(X, \mathcal{B}, \mu)$, where X is a set, $\mathcal{B}$ is a σ-algebra of subsets of X, and $\mu : \mathcal{B} \to [0, +\infty]$ is a countably additive unsigned measure on $\mathcal{B}$. If the measure $\mu(X)$ of the total space is one, then the measure space becomes a *probability space*. If a non-negative function $f : X \to [0, +\infty]$ is $\mathcal{B}$-*measurable* (or *measurable* for short), one can then form the integral $\int_X f \, d\mu \in [0, +\infty]$ by the usual abstract measure-theoretic construction (as discussed for instance in [**Ta2011**, §1.4]).

A measure space is *complete* if every subset of a null set (i.e. a measurable set of measure zero) is also a null set. Not all measure spaces are complete, but one can always form the *completion* $(X, \overline{\mathcal{B}}, \mu)$ of a measure space $(X, \mathcal{B}, \mu)$ by enlarging the σ-algebra $\mathcal{B}$ to the space of all sets which are equal to a measurable set outside of a null set, and extending the measure μ appropriately.

Given two (σ-finite) measure spaces $(X, \mathcal{B}_X, \mu_X)$ and $(Y, \mathcal{B}_Y, \mu_Y)$, one can form the *product space* $(X \times Y, \mathcal{B}_X \times \mathcal{B}_Y, \mu_X \times \mu_Y)$. This is a measure space whose domain is the Cartesian product $X \times Y$, the σ-algebra $\mathcal{B}_X \times \mathcal{B}_Y$ is generated by the "rectangles" $A \times B$ with $A \in \mathcal{B}_X$, $B \in \mathcal{B}_Y$, and the measure $\mu_X \times \mu_Y$ is the unique measure on $\mathcal{B}_X \times \mathcal{B}_Y$ obeying the identity

$$\mu_X \times \mu_Y(A \times B) = \mu_X(A)\mu_Y(B).$$

See for instance [**Ta2011**, §1.7] for a formal construction of product measure[2]. One of the fundamental theorems concerning product measure is *Tonelli's theorem* (which is basically the unsigned version of the more well-known *Fubini theorem*), which asserts that if $f : X \times Y \to [0, +\infty]$ is $\mathcal{B}_X \times \mathcal{B}_Y$ measurable, then the integral expressions

$$\int_X (\int_Y f(x, y) \, d\mu_Y(y)) \, d\mu_X(x),$$

$$\int_Y (\int_X f(x, y) \, d\mu_X(x)) \, d\mu_Y(y)$$

and

$$\int_{X \times Y} f(x, y) \, d\mu_{X \times Y}(x, y)$$

all exist (thus all integrands are almost-everywhere well-defined and measurable with respect to the appropriate σ-algebras), and are all equal to each other; see e.g. [**Ta2011**, Theorem 1.7.15].

[2]There are technical difficulties with the theory when X or Y is not σ-finite, but in these notes we will only be dealing with probability spaces, which are clearly σ-finite, so this difficulty will not concern us.

Any finite non-empty set V can be turned into a probability space $(V, 2^V, \mu_V)$ by endowing it with the discrete σ-algebra $2^V := \{A : A \subset V\}$ of all subsets of V, and the normalised counting measure

$$\mu(A) := \frac{|A|}{|V|},$$

where $|A|$ denotes the cardinality of A. In this discrete setting, the probability space is automatically complete, and every function $f : V \to [0, +\infty]$ is measurable, with the integral simply being the average:

$$\int_V f \, d\mu_V = \frac{1}{|V|} \sum_{v \in V} f(v).$$

Of course, Tonelli's theorem is obvious for these discrete spaces; the deeper content of that theorem is only apparent at the level of continuous measure spaces.

Among other things, this probability space structure on finite sets can be used to describe various statistics of dense graphs. Recall that a graph $G = (V, E)$ is a finite vertex set V, together with a set of edges E, which we will think of as a symmetric subset[3] of the Cartesian product $V \times V$. Then, if V is non-empty, and ignoring some minor errors coming from the diagonal V^Δ, the *edge density* of the graph is essentially

$$e(G) := \mu_{V \times V}(E) = \int_{V \times V} 1_E(v, w) \, d\mu_{V \times V}(v, w),$$

the *triangle density* of the graph is basically

$$t(G) := \int_{V \times V \times V} 1_E(u, v) 1_E(v, w) 1_E(w, u) \, d\mu_{V \times V \times V}(u, v, w),$$

and so forth.

In [**RuSz1978**], Ruzsa and Szemerédi established the *triangle removal lemma* concerning triangle densities, which informally asserts that a graph with few triangles can be made completely triangle-free by removing a small number of edges:

Lemma 5.2.1 (Triangle removal lemma)**.** *Let $G = (V, E)$ be a graph on a non-empty finite set V, such that $t(G) \leq \delta$ for some $\delta > 0$. Then there exists a subgraph $G' = (V, E')$ of G with $t(G') = 0$, such that $e(G \backslash G') = o_{\delta \to 0}(1)$, where $o_{\delta \to 0}(1)$ denotes a quantity bounded by $c(\delta)$ for some function $c(\delta)$ of δ that goes to zero as $\delta \to 0$.*

[3] If one wishes, one can prohibit loops in E, so that E is disjoint from the diagonal $V^\Delta := \{(v, v) : v \in V\}$ of $V \times V$, but this will not make much difference for the discussion below.

The original proof of the triangle removal lemma was a "finitary" one, and proceeded via the *Szemerédi regularity lemma* [**Sz1978**]. It has a number of consequences; for instance, as already noted in that paper, the triangle removal lemma implies as a corollary *Roth's theorem* [**Ro1953**] that subsets of $\mathbf{Z}$ of positive upper density contain infinitely many arithmetic progressions of length three.

It is however also possible to establish this lemma by infinitary means. There are at least three basic approaches for this. One is via a correspondence principle between questions about dense finite graphs, and questions about exchangeable random infinite graphs, as was pursued in [**Ta2007**], [**Ta2010b**, §2.3]. A second (closely related to the first) is to use the machinery of graph limits, as developed in [**LoSz2006**], [**BoChLoSoVe2008**]. The third is via nonstandard analysis (or equivalently, by using ultraproducts), as was pursued in [**ElSz2012**]. These three approaches differ in the technical details of their execution, but the net effect of all of these approaches is broadly the same, in that they both convert statements about large dense graphs (such as the triangle removal lemma) to measure theoretic statements on infinitary measure spaces. (This is analogous to how the Furstenberg correspondence principle converts combinatorial statements about dense sets of integers into ergodic-theoretic statements on measure-preserving systems.)

In this section, we will illustrate the nonstandard analysis approach of [**ElSz2012**] by providing a nonstandard proof of the triangle removal lemma. The main technical tool used here (besides the basic machinery of nonstandard analysis) is that of *Loeb measure* [**Lo1975**], which gives a probability space structure $(V, \mathcal{B}_V, \mu_V)$ to *nonstandard* finite non-empty sets $V = \prod_{n \to p} V_n$ that is an infinitary analogue of the discrete probability space structures $V = (V, 2^V, \mu_V)$ one has on standard finite non-empty sets. The nonstandard analogue of quantities such as triangle densities then become the integrals of various nonstandard functions with respect to Loeb measure. With this approach, the epsilons and deltas that are so prevalent in the finitary approach to these subjects disappear almost completely; but to compensate for this, one now must pay much more attention to questions of measurability, which were automatic in the finitary setting but now require some care in the infinitary one.

The nonstandard analysis approaches are also related to the regularity lemma approach; see [**Ta2011d**, §4.4] for a proof of the regularity lemma using Loeb measure.

As usual, the nonstandard approach offers a complexity tradeoff: there is more effort expended in building the foundational mathematical structures of the argument (in this case, ultraproducts and Loeb measure), but

once these foundations are completed, the actual arguments are shorter than their finitary counterparts. In the case of the triangle removal lemma, this tradeoff does not lead to a particularly significant reduction in complexity (and arguably leads in fact to an increase in the length of the arguments, when written out in full), but the gain becomes more apparent when proving more complicated results, such as the hypergraph removal lemma, in which the initial investment in foundations leads to a greater savings in net complexity, as can be seen in [**ElSz2012**].

5.2.1. Loeb measure. We use the usual setup of nonstandard analysis (as reviewed for instance in [**Ta2011d**, §4.4]). Thus, we will need a non-principal *Ultrafilter*ultrafilter $p \in \beta\mathbf{N}\backslash\mathbf{N}$ on the natural numbers $\mathbf{N}$. A statement $P(n)$ pertaining to a natural number n is said to hold *for n sufficiently close to p* if the set of n for which $P(n)$ holds lies in the ultrafilter p. Given a sequence X_n of (standard) spaces X_n, the *Ultraproduct*ultraproduct $\prod_{n\to p} X_n$ is the space of all *ultralimits* $\lim_{n\to p} x_n$ with $x_n \in X_n$, with two ultralimits $\lim_{n\to p} x_n, \lim_{n\to p} y_n$ considered equal if and only if $x_n = y_n$ for all n sufficiently close to p.

Now consider a nonstandard finite non-empty set V, i.e. an ultraproduct $V = \prod_{n\to p} V_n$ of standard finite non-empty sets V_n. Define an *internal subset* of V to be a subset of V of the form $A = \prod_{n\to p} A_n$, where each A_n is a subset of V_n. It is easy to see that the collection $\mathcal{A}_V$ of all internal subsets of V is a boolean algebra. In general, though, $\mathcal{A}_V$ will not be a σ-algebra. For instance, suppose that the V_n are the standard discrete intervals $V_n := [1,n] := \{i \in \mathbf{N} : i \leq n\}$, then V is the non-standard discrete interval $V = [1,N] := \{i \in {}^*\mathbf{N} : i \leq N\}$, where N is the unbounded nonstandard natural number $N := \lim_{n\to p} n$. For any standard integer m, the subinterval $[1, N/m]$ is an internal subset of V; but the intersection

$$[1, o(N)] := \bigcap_{m\in\mathbf{N}} [1, N/m] = \{i \in {}^*\mathbf{N} : i = o(N)\}$$

is not an internal subset of V. (This can be seen, for instance, by noting that all non-empty internal subsets of $[1,N]$ have a maximal element, whereas $[1, o(N)]$ does not.)

Given any internal subset $A = \prod_{n\to p} A_n$ of V, we can define the cardinality $|A|$ of A, which is the nonstandard natural number $|A| := \lim_{n\to p} |A_n|$. We then have the nonstandard density $\frac{|A|}{|V|}$, which is a nonstandard real number between 0 and 1. By the Bolzano-Weierstrass theorem, every this bounded nonstandard real number $\frac{|A|}{|V|}$ has a unique *standard part* $\mathrm{st}(\frac{|A|}{|V|})$, which is a standard real number in $[0,1]$ such that

$$\frac{|A|}{|V|} = \mathrm{st}(\frac{|A|}{|V|}) + o(1),$$

where $o(1)$ denotes a nonstandard infinitesimal (i.e. a nonstandard number which is smaller in magnitude than any standard $\varepsilon > 0$).

In [**Lo1975**], Loeb observed that this standard density can be extended to a complete probability measure:

Theorem 5.2.2 (Construction of Loeb measure)**.** *Let V be a nonstandard finite non-empty set. Then there exists a complete probability space $(V, \mathcal{L}_V, \mu_V)$, with the following properties:*

- *(Internal sets are Loeb measurable) If A is an internal subset of V, then $A \in \mathcal{L}_V$ and*
$$\mu_V(A) = st(\frac{|A|}{|V|}).$$
- *(Loeb measurable sets are almost internal) If E is a subset of V, then E is Loeb measurable if and only if, for every standard $\varepsilon > 0$, there exists internal subsets $A, B_1, B_2, \ldots$ of V such that*
$$E \Delta A \subset \bigcup_{n=1}^{\infty} B_n$$
and
$$\sum_{n=1}^{\infty} \mu_V(B_n) \leq \varepsilon.$$

Proof. The map $\mu_V : A \mapsto \mathrm{st}(\frac{|A|}{|V|})$ is a finitely additive probability measure on $\mathcal{A}_V$. We claim that this map μ_V is in fact a *pre-measure* on $\mathcal{A}_V$, thus one has

$$\mu_V(A) = \sum_{n=1}^{\infty} \mu_V(A_n) \tag{5.2}$$

whenever A is an internal set that is partitioned into a disjoint sequence of internal sets A_n. But the countable sequence of sets $A \backslash (A_1 \cup \ldots A_n)$ are internal, and have empty intersection, so by the *countable saturation* property of ultraproducts (see e.g. [**Ta2011d**, §4.4]), one of the $A \backslash (A_1 \cup \ldots \cup A_n)$ must be empty. The pre-measure property (5.2) then follows from the finite additivity of μ_V.

Invoking the *Hahn-Kolmogorov extension theorem* (see e.g. [**Ta2011**, Theorem 1.7.8]), we conclude that μ_V extends to a countably additive probability measure on the σ-algebra $\langle \mathcal{A}_V \rangle$ generated by the internal sets. This measure need not be complete, but we can then pass to the completion $\mathcal{L}_V := \overline{\langle \mathcal{A}_V \rangle}$ of that σ-algebra. This probability space certainly obeys the first property. The "only if" portion of second property asserts that all Loeb measurable sets differ from an internal set by sets of arbitrarily small

outer measure, but this is easily seen since the space of all sets that have this property is easily verified to be a complete σ-algebra that contain the algebra of internal sets. The "if" portion follows easily from the fact that $\mathcal{L}_V$ is a complete σ-algebra containing the internal sets. (These facts are very similar to the more familiar facts that a bounded subset of a Euclidean space is Lebesgue measurable if and only if it differs from an elementary set by a set of arbitrarily small outer measure.) □

Now we turn to the analogue of Tonelli's theorem for Loeb measure, which will be a fundamental tool when it comes to prove the triangle removal lemma. Let V, W be two nonstandard finite non-empty sets, then $V \times W$ is also a nonstandard finite non-empty set. We then have three Loeb probability spaces

$$\begin{aligned} &(V, \mathcal{L}_V, \mu_V), \\ &(W, \mathcal{L}_W, \mu_W), \\ &(V \times W, \mathcal{L}_{V\times W}, \mu_{V\times W}), \end{aligned} \tag{5.3}$$

and we also have the product space

$$(V \times W, \mathcal{L}_V \times \mathcal{L}_W, \mu_V \times \mu_W). \tag{5.4}$$

It is then natural to ask how the two probability spaces (5.3) and (5.4) are related. There is one easy relationship, which shows that (5.3) extends (5.4):

Exercise 5.2.1. Show that (5.3) is a refinement of (5.4), thus $\mathcal{L}_V \times \mathcal{L}_W$, and $\mu_{V\times W}$ extends $\mu_V \times \mu_W$. (*Hint:* first recall why the product of Lebesgue measurable sets is Lebesgue measurable, and mimic that proof to show that the product of a $\mathcal{L}_V$-measurable set and a $\mathcal{L}_W$-measurable set is $\mathcal{L}_{V\times W}$-measurable, and that the two measures $\mu_{V\times W}$ and $\mu_V \times \mu_W$ agree in this case.)

In the converse direction, (5.3) enjoys the type of Tonelli theorem that (5.4) does:

Theorem 5.2.3 (Tonelli theorem for Loeb measure)**.** *Let V, W be two nonstandard finite non-empty sets, and let $f : V \times W \to [0, +\infty]$ be an unsigned $\mathcal{L}_{V\times W}$-measurable function. Then the expressions*

$$\int_V (\int_W f(v,w)\ d\mu_W(w))\ d\mu_V(v) \tag{5.5}$$

$$\int_W (\int_V f(v,w)\ d\mu_W(w))\ d\mu_V(v) \tag{5.6}$$

and

$$\int_{V\times W} f(v,w)\ d\mu_{V\times W}(v,w) \tag{5.7}$$

are well-defined (thus all integrands are almost everywhere well-defined and appropriately measurable) and equal to each other.

Proof. By the monotone convergence theorem it suffices to verify this when f is a simple function; by linearity we may then take f to be an indicator function $f = 1_E$. Using Theorem 5.2.2 and an approximation argument (and many further applications of monotone convergence) we may assume without loss of generality that E is an internal set. We then have

$$\int_{V\times W} f(v,w)\ d\mu_{V\times W}(v,w) = \mathrm{st}(\frac{|E|}{|V||W|})$$

and for every $v \in V$, we have

$$\int_{W} f(v,w)\ d\mu_{W}(w) = \mathrm{st}(\frac{|E_v|}{|W|}),$$

where E_v is the internal set

$$E_v := \{w \in W : (v,w) \in E\}.$$

Let n be a standard natural number, then we can partition V into the internal sets $V = V_1 \cup \ldots \cup V_n$, where

$$V_i := \{v \in V : \frac{i-1}{n} < \frac{|E_v|}{|W|} \leq \frac{i}{n}\}.$$

On each V_i, we have

$$\int_{W} f(v,w)\ d\mu_{W}(w) = \frac{i}{n} + O(\frac{1}{n}) \tag{5.8}$$

and

$$\frac{|E_v|}{|W|} = \frac{i}{n} + O(\frac{1}{n}). \tag{5.9}$$

From (5.8), we see that the upper and lower integrals of $\int_W f(v,w)\ d\mu_W(w)$ are both of the form

$$\sum_{i=1}^{n} \frac{i}{n}\frac{|V_i|}{|V|} + O(\frac{1}{n}).$$

Meanwhile, using the nonstandard double counting identity

$$\frac{1}{|V|}\sum_{v\in V} \frac{|E_v|}{|W|} = \frac{|E|}{|V||W|}$$

(where all arithmetic operations are interpreted in the nonstandard sense, of course) and (5.9), we see that

$$\frac{|E|}{|V||W|} = \sum_{i=1}^{n} \frac{i}{n} \frac{|V_i|}{|V|} + O(\frac{1}{n}).$$

Thus we see that the upper and lower integrals of $\int_W f(v,w)\ d\mu_W(w)$ are equal to $\frac{|E|}{|V||W|} + O(\frac{1}{n})$ for every standard n. Sending n to infinity, we conclude that $\int_W f(v,w)\ d\mu_W(w)$ is measurable, and that

$$\int_V (\int_W f(v,w)\ d\mu_W(w))\ d\mu_V(v) = \mathrm{st}(\frac{|E|}{|V||W|})$$

showing that (5.5) and (5.7) are well-defined and equal. A similar argument holds for (5.6) and (5.7), and the claim follows. □

Remark 5.2.4. It is well known that the product of two Lebesgue measure spaces $\mathbf{R}^n, \mathbf{R}^m$, upon completion, becomes the Lebesgue measure space on $\mathbf{R}^{n+m}$. Drawing the analogy between Loeb measure and Lebesgue measure, it is then natural to ask whether (5.3) is simply the completion of (5.4). But while (5.3) certainly contains the completion of (5.4), it is a significantly larger space in general. Indeed, suppose $V = \prod_{n\to p} V_n$, $W = \prod_{n\to p} W_n$, where the cardinality of V_n, W_n goes to infinity at some reasonable rate, e.g. $|V_n|, |W_n| \geq n$ for all n. For each n, let E_n be a random subset of $V_n \times W_n$, with each element of $V_n \times W_n$ having an independent probability of $1/2$ of lying in E_n. Then, as is well known, the sequence of sets E_n is almost surely *asymptotically regular* in the sense that almost surely, we have the bound

$$\sup_{A_n \subset V_n, B_n \subset W_n} \frac{||E_n \cap (A_n \times B_n)| - \frac{1}{2}|A_n||B_n||}{|V_n||W_n|} \to 0$$

as $n \to \infty$. Let us condition on the event that this asymptotic regularity holds. Taking ultralimits, we conclude that the internal set $E := \prod_{n\to p} E_n$ obeys the property

$$\mu_{V\times W}(E \cap (A \times B)) = \frac{1}{2}\mu_{V\times W}(A \times B)$$

for all internal $A \subset V, B \subset W$; in particular, E has Loeb measure $1/2$. Using Theorem 5.2.2 we conclude that

$$\mu_{V\times W}(E \cap F) = \frac{1}{2}\mu_{V\times W}(F)$$

for all $\mathcal{L}_V \times \mathcal{L}_W$-measurable F, which implies in particular that E cannot be $\mathcal{L}_V \times \mathcal{L}_W$-measurable. (Indeed, $1_E - \frac{1}{2}$ is "anti-measurable" in the sense that it is orthogonal to all functions in $L^2(\mathcal{L}_V \times \mathcal{L}_W)$; or equivalently, we have the conditional expectation formula $\mathbf{E}(1_E|\mathcal{L}_V \times \mathcal{L}_W) = \frac{1}{2}$ almost everywhere.)

Intuitively, a $\mathcal{L}_V \times \mathcal{L}_W$-measurable set corresponds to a subset of $V \times W$ that is of "almost bounded complexity", in that it can be approximated by a bounded boolean combination of Cartesian products. In contrast, $\mathcal{L}_{V\times W}$-measurable sets (such as the set E given above) have no bound on their complexity.

5.2.2. The triangle removal lemma. Now we can prove the triangle removal lemma, Lemma 5.2.1. We will deduce it from the following nonstandard (and tripartite) counterpart (a special case of a result first established in [**Ta2007**]):

Lemma 5.2.5 (Nonstandard triangle removal lemma)**.** *Let V be a nonstandard finite non-empty set, and let $E_{12}, E_{23}, E_{31} \subset V \times V$ be Loeb-measurable subsets of $V \times V$ which are almost triangle-free in the sense that*

$$\int_{V\times V\times V} 1_{E_{12}}(u,v)1_{E_{23}}(v,w)1_{E_{31}}(w,u)\ d\mu_{V\times V\times V}(u,v,w) = 0. \tag{5.10}$$

Then for any standard $\varepsilon > 0$, there exists a internal subsets $F_{ij} \subset V \times V$ for $ij = 12, 23, 31$ with $\mu_{V\times V}(E_{ij}\backslash F_{ij}) < \varepsilon$, which are completely triangle-free in the sense that

$$1_{F_{12}}(u,v)1_{F_{23}}(v,w)1_{F_{31}}(w,u) = 0 \tag{5.11}$$

for all $u, v, w \in V$.

Let us first see why Lemma 5.2.5 implies Lemma 5.2.1. We use the usual "compactness and contradiction" argument. Suppose for contradiction that Lemma 5.2.1 failed. Carefully negating the quantifiers, we can find a (standard) $\varepsilon > 0$, and a sequence $G_n = (V_n, E_n)$ of graphs with $t(G_n) \leq 1/n$, such that for each n, there does *not* exist a subgraph $G'_n = (V_n, E'_n)$ of n with $|E_n\backslash E'_n| \leq \varepsilon |V_n|^2$ with $t(G'_n) = 0$. Clearly we may assume the V_n are non-empty.

We form the ultraproduct $G = (V, E)$ of the G_n, thus $V = \prod_{n\to p} V_n$ and $E = \prod_{n\to p} E_n$. By construction, E is a symmetric internal subset of $V \times V$ and we have

$$\int_{V\times V\times V} 1_E(u,v)1_E(v,w)1_E(w,u)\ d\mu_{V\times V\times V}(u,v,w) = \mathrm{st} \lim_{n\to p} t(G_n) = 0.$$

Thus, by Lemma 5.2.5, we may find internal subsets F_{12}, F_{23}, F_{31} of $V \times V$ with $\mu_{V\times V}(E\backslash F_{ij}) < \varepsilon/6$ (say) for $ij = 12, 23, 31$ such that (5.11) holds for all $u, v, w \in V$. By letting E' be the intersection of all E with all the F_{ij} and their reflections, we see that E' is a symmetric internal subset of E with $\mu_{V\times V}(E\backslash E') < \varepsilon$, and we still have

$$1_{E'}(u,v)1_{E'}(v,w)1_{E'}(w,u) = 0$$

for all $u, v, w \in V$. If we write $E' = \lim_{n \to p} E'_n$ for some sets E'_n, then for n sufficiently close to p, one has E'_n a symmetric subset of E_n with

$$\mu_{V_n \times V_n}(E_n \backslash E'_n) < \varepsilon$$

and

$$1_{E'_n}(u,v) 1_{E'_n}(v,w) 1_{E'_n}(w,u) = 0.$$

If we then set $G'_n := (V_n, E_n)$, we thus have $|E_n \backslash E'_n| \leq \varepsilon |V_n|^2$ and $t(G'_n) = 0$, which contradicts the construction of G_n by taking n sufficiently large.

Now we prove Lemma 5.2.5. The idea (similar to that used to prove the Furstenberg recurrence theorem, as discussed for instance in [**Ta2009**, §2.15]) is to first prove the lemma for very simple examples of sets E_{ij}, and then work one's way towards the general case. Readers who are familiar with the traditional proof of the triangle removal lemma using the regularity lemma will see strong similarities between that argument and the one given here (and, on some level, they are essentially the same argument).

To begin with, we suppose first that the E_{ij} are all *elementary sets*, in the sense that they are finite boolean combinations of products of internal sets. (At the finitary level, this corresponds to graphs that are bounded combinations of bipartite graphs.) This implies that there is an internal partition $V = V_1 \cup \ldots \cup V_n$ of the vertex set V, such that each E_{ij} is the union of some of the $V_a \times V_b$.

Let F_{ij} be the union of all the $V_a \times V_b$ in E_{ij} for which V_a and V_b have positive Loeb measure; then $\mu_{V \times V}(E_{ij} \backslash F_{ij}) = 0$. We claim that (5.11) holds for all $u, v, w \in V$, which gives Theorem 5.2.5 in this case. Indeed, if $u \in V_a, v \in V_b, w \in V_c$ were such that (5.11) failed, then E_{12} would contain $V_a \times V_b$, E_{23} would contain $V_b \times V_c$, and E_{31} would contain $V_c \times V_a$. The integrand in (5.10) is then equal to 1 on $V_a \times V_b \times V_c$, which has Loeb measure $\mu_V(V_a)\mu_V(V_b)\mu_V(V_c)$ which is non-zero, contradicting (5.10). This gives Theorem 5.2.5 in the elementary set case.

Next, we increase the level of generality by assuming that the E_{ij} are all $\overline{\mathcal{L}_V \times \mathcal{L}_V}$-measurable. (The finitary equivalent of this is a little difficult to pin down; roughly speaking, it is dealing with graphs that are not quite bounded combinations of bounded graphs, but can be well approximated by such bounded combinations; a good example is the *half-graph*, which is a bipartite graph between two copies of $\{1, \ldots, N\}$, which joins an edge between the first copy of i and the second copy of j iff $i < j$.) Then each E_{ij} can be approximated to within an error of $\varepsilon/3$ in $\mu_{V \times V}$ by elementary sets. In particular, we can find a finite partition $V = V_1 \cup \ldots \cup V_n$ of V, and sets E'_{ij} that are unions of some of the $V_a \times V_b$, such that $\mu_{V \times V}(E_{ij} \Delta E'_{ij}) < \varepsilon/3$.

Let F_{ij} be the union of all the $V_a \times V_b$ contained in E'_{ij} such that V_a, V_b have positive Loeb measure, and such that

$$\mu_{V\times V}(E_{ij} \cap (V_a \times V_b)) > \frac{2}{3}\mu_{V\times V}(V_a \times V_b).$$

Then the F_{ij} are internal subsets of $V \times V$, and $\mu_{V\times V}(E_{ij}\backslash F_{ij}) < \varepsilon$.

We now claim that the F_{ij} obey (5.11) for all u, v, w, which gives Theorem 5.2.5 in this case. Indeed, if $u \in V_a, v \in V_b, w \in V_c$ were such that (5.11) failed, then E_{12} occupies more than $\frac{2}{3}$ of $V_a \times V_b$, and thus

$$\int_{V_a\times V_b\times V_c} 1_{E_{12}}(u,v)\ d\mu_{V\times V\times V}(u,v,w) > \frac{2}{3}\mu_{V\times V\times V}(V_a \times V_b \times V_c).$$

Similarly for $1_{E_{23}}(v,w)$ and $1_{E_{31}}(w,u)$. From the inclusion-exclusion formula, we conclude that

$$\int_{V_a\times V_b\times V_c} 1_{E_{12}}(u,v)1_{E_{23}}(v,w)1_{E_{31}}(w,u)\ d\mu_{V\times V\times V}(u,v,w) > 0,$$

contradicting (5.10), and the claim follows.

Finally, we turn to the general case, when the E_{ij} are merely $\mathcal{L}_{V\times V}$-measurable. Here, we split

$$1_{E_{ij}} = f_{ij} + g_{ij}$$

where $f_{ij} := \mathbf{E}(1_{E_{ij}}|\overline{\mathcal{L}_V \times \mathcal{L}_V})$ is the conditional expectation of $1_{E_{ij}}$ onto $\overline{\mathcal{L}_V \times \mathcal{L}_V}$, and $g_{ij} := 1_{E_{ij}} - f_{ij}$ is the remainder. We observe that each $g_{ij}(u,v)$ is orthogonal to any tensor product $f(u)g(v)$ with f, g bounded and $\mathcal{L}_V$-measurable. From this and Tonelli's theorem for Loeb measure (Theorem 5.2.3) we conclude that each of the g_{ij} make a zero contribution to (5.10), and thus

$$\int_{V\times V\times V} f_{12}(u,v)f_{23}(v,w)f_{31}(w,u)\ d\mu_{V\times V\times V}(u,v,w) = 0.$$

Now let $E'_{ij} := \{(u,v) \in V \times V : f_{ij}(u,v) \geq \varepsilon/2\}$, then the E'_{ij} are $\overline{\mathcal{L}_V \times \mathcal{L}_V}$-measurable, and we have

$$\int_{V\times V\times V} 1_{E'_{12}}(u,v)1_{E'_{23}}(v,w)1_{E'_{31}}(w,u)\ d\mu_{V\times V\times V}(u,v,w) = 0.$$

Also, we have

$$\begin{aligned}\mu_{V\times V}(E_{ij}\backslash E'_{ij}) &= \int_{V\times V} 1_{E_{ij}}(1 - 1_{E'_{ij}}) \\ &= \int_{V\times V} f_{ij}(1 - 1_{E'_{ij}}) \\ &\leq \varepsilon/2.\end{aligned}$$

Applying the already established cases of Theorem 5.2.5, we can find internal sets F_{ij} obeying (5.11) with $\mu_{V\times V}(E'_{ij}\backslash F_{ij}) < \varepsilon/2$, and hence $\mu_{V\times V}(E_{ij}\backslash F_{ij}) < \varepsilon$, and Theorem 5.2.5 follows.

Remark 5.2.6. The full hypergraph removal lemma can be proven using similar techniques, but with a longer tower of generalisations than the three cases given here; see [**Ta2007**] or [**ElSz2012**].

Chapter 6

Partial differential equations

6.1. The limiting absorption principle

Perhaps the most fundamental differential operator on Euclidean space $\mathbf{R}^d$ is the *Laplacian*

$$\Delta := \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2}.$$

The Laplacian is a linear translation-invariant operator, and as such is necessarily diagonalised by the Fourier transform

$$\hat{f}(\xi) := \int_{\mathbf{R}^d} f(x) e^{-2\pi i x\cdot\xi}\ dx.$$

Indeed, we have

$$\widehat{\Delta f}(\xi) = -4\pi^2|\xi|^2 \hat{f}(\xi)$$

for any suitably nice function f (e.g. in the Schwartz class; alternatively, one can work in very rough classes, such as the space of tempered distributions, provided of course that one is willing to interpret all operators in a distributional or weak sense).

Because of this explicit diagonalisation, it is a straightforward manner to define spectral multipliers $m(-\Delta)$ of the Laplacian for any (measurable, polynomial growth) function $m : [0, +\infty) \to \mathbf{C}$, by the formula

$$\widehat{m(-\Delta)f}(\xi) := m(4\pi^2|\xi|^2)\hat{f}(\xi).$$

(The presence of the minus sign in front of the Laplacian has some minor technical advantages, as it makes $-\Delta$ positive semi-definite. One can also

define spectral multipliers more abstractly from general *functional calculus*, after establishing that the Laplacian is essentially self-adjoint.) Many of these multipliers are of importance in PDE and analysis, such as the *fractional derivative* operators $(-\Delta)^{s/2}$, the *heat propagators* $e^{t\Delta}$, the (free) *Schrödinger propagators* $e^{it\Delta}$, the *wave propagators* $e^{\pm it\sqrt{-\Delta}}$ (or $\cos(t\sqrt{-\Delta})$ and $\frac{\sin(t\sqrt{-\Delta})}{\sqrt{-\Delta}}$, depending on one's conventions), the spectral projections $1_I(\sqrt{-\Delta})$, the *Bochner-Riesz summation operators* $(1 + \frac{\Delta}{4\pi^2 R^2})^\delta_+$, or the *resolvents* $R(z) := (-\Delta - z)^{-1}$.

Each of these families of multipliers are related to the others, by means of various integral transforms (and also, in some cases, by analytic continuation). For instance:

(1) Using the *Laplace transform*, one can express (sufficiently smooth) multipliers in terms of heat operators. For instance, using the identity
$$\lambda^{s/2} = \frac{1}{\Gamma(-s/2)} \int_0^\infty t^{-1-s/2} e^{-t\lambda}\ dt$$
(using analytic continuation if necessary to make the right-hand side well-defined), with Γ being the *Gamma function*, we can write the fractional derivative operators in terms of heat kernels:
$$(-\Delta)^{s/2} = \frac{1}{\Gamma(-s/2)} \int_0^\infty t^{-1-s/2} e^{t\Delta}\ dt. \tag{6.1}$$

(2) Using analytic continuation, one can connect heat operators $e^{t\Delta}$ to Schrödinger operators $e^{it\Delta}$, a process also known as *Wick rotation*. Analytic continuation is a notoriously unstable process, and so it is difficult to use analytic continuation to obtain any quantitative estimates on (say) Schrödinger operators from their heat counterparts; however, this procedure can be useful for propagating *identities* from one family to another. For instance, one can derive the fundamental solution for the Schrödinger equation from the fundamental solution for the heat equation by this method.

(3) Using the *Fourier inversion formula*, one can write general multipliers as integral combinations of Schrödinger or wave propagators; for instance, if z lies in the upper half plane $\mathbf{H} := \{z \in \mathbf{C} : \mathrm{Im}\, z > 0\}$, one has
$$\frac{1}{x - z} = i \int_0^\infty e^{-itx} e^{itz}\ dt$$
for any real number x, and thus we can write resolvents in terms of Schrödinger propagators:
$$R(z) = i \int_0^\infty e^{it\Delta} e^{itz}\ dt. \tag{6.2}$$

In a similar vein, if $k \in \mathbf{H}$, then

$$\frac{1}{x^2 - k^2} = \frac{i}{k} \int_0^\infty \cos(tx) e^{ikt}\ dt$$

for any $x > 0$, so one can also write resolvents in terms of wave propagators:

$$R(k^2) = \frac{i}{k} \int_0^\infty \cos(t\sqrt{-\Delta}) e^{ikt}\ dt. \tag{6.3}$$

(4) Using the *Cauchy integral formula*, one can express (sufficiently holomorphic) multipliers in terms of resolvents (or limits of resolvents). For instance, if $t > 0$, then from the Cauchy integral formula (and *Jordan's lemma*) one has

$$e^{itx} = \frac{1}{2\pi i} \lim_{\varepsilon \to 0^+} \int_{\mathbf{R}} \frac{e^{ity}}{y - x + i\varepsilon}\ dy$$

for any $x \in \mathbf{R}$, and so one can (formally, at least) write Schrödinger propagators in terms of resolvents:

$$e^{-it\Delta} = -\frac{1}{2\pi i} \lim_{\varepsilon \to 0^+} \int_{\mathbf{R}} e^{ity} R(y + i\varepsilon)\ dy. \tag{6.4}$$

(5) The imaginary part of $\frac{1}{\pi} \frac{1}{x-(y+i\varepsilon)}$ is the *Poisson kernel* $\frac{\varepsilon}{\pi} \frac{1}{(y-x)^2+\varepsilon^2}$, which is an approximation to the identity. As a consequence, for any reasonable function $m(x)$, one has (formally, at least)

$$m(x) = \lim_{\varepsilon \to 0^+} \frac{1}{\pi} \int_{\mathbf{R}} (\operatorname{Im} \frac{1}{x - (y + i\varepsilon)}) m(y)\ dy$$

which leads (again formally) to the ability to express arbitrary multipliers in terms of imaginary (or skew-adjoint) parts of resolvents:

$$m(-\Delta) = \lim_{\varepsilon \to 0^+} \frac{1}{\pi} \int_{\mathbf{R}} (\operatorname{Im} R(y + i\varepsilon)) m(y)\ dy. \tag{6.5}$$

Among other things, this type of formula (with $-\Delta$ replaced by a more general self-adjoint operator) is used in the resolvent-based approach to the spectral theorem (by using the limiting imaginary part of resolvents to build spectral measure). Note that one can also express $\operatorname{Im} R(y + i\varepsilon)$ as $\frac{1}{2i}(R(y + i\varepsilon) - R(y - i\varepsilon))$.

Remark 6.1.1. The ability of heat operators, Schrödinger propagators, wave propagators, or resolvents to generate other spectral multipliers can be viewed as a sort of manifestation of the *Stone-Weierstrass theorem* (though with the caveat that the spectrum of the Laplacian is non-compact and so

the Stone-Weierstrass theorem does not directly apply). Indeed, observe the **-algebra* type identities

$$\begin{aligned} e^{s\Delta} e^{t\Delta} &= e^{(s+t)\Delta}; \\ (e^{s\Delta})^* &= e^{s\Delta}; \\ e^{is\Delta} e^{it\Delta} &= e^{i(s+t)\Delta}; \\ (e^{is\Delta})^* &= e^{-is\Delta}; \\ e^{is\sqrt{-\Delta}} e^{it\sqrt{-\Delta}} &= e^{i(s+t)\sqrt{-\Delta}}; \\ (e^{is\sqrt{-\Delta}})^* &= e^{-is\sqrt{-\Delta}}; \\ R(z)R(w) &= \frac{R(w) - R(z)}{z - w}; \\ R(z)^* &= R(\overline{z}). \end{aligned}$$

Because of these relationships, it is possible (in principle, at least), to leverage one's understanding one family of spectral multipliers to gain control on another family of multipliers. For instance, the fact that the heat operators $e^{t\Delta}$ have non-negative kernel (a fact which can be seen from the maximum principle, or from the Brownian motion interpretation of the heat kernels) implies (by (6.1)) that the fractional integral operators $(-\Delta)^{-s/2}$ for $s > 0$ also have non-negative kernel. Or, the fact that the wave equation enjoys finite speed of propagation (and hence that the wave propagators $\cos(t\sqrt{-\Delta})$ have distributional convolution kernel localised to the ball of radius $|t|$ centred at the origin), can be used (by (6.3)) to show that the resolvents $R(k^2)$ have a convolution kernel that is essentially localised to the ball of radius $O(1/|\operatorname{Im}(k)|)$ around the origin.

In this section, we will continue this theme by using the *resolvents* $R(z) = (-\Delta - z)^{-1}$ to control other spectral multipliers. These resolvents are well-defined whenever z lies outside of the spectrum $[0, +\infty)$ of the operator $-\Delta$. In the model three-dimensional[1] case $d = 3$, they can be defined explicitly by the formula

$$R(k^2) f(x) = \int_{\mathbf{R}^3} \frac{e^{ik|x-y|}}{4\pi|x-y|} f(y)\, dy$$

whenever k lives in the upper half-plane $\{k \in \mathbf{C} : \operatorname{Im}(k) > 0\}$, ensuring the absolute convergence of the integral for test functions f. It is an instructive exercise to verify that this resolvent indeed inverts the operator $-\Delta - k^2$, either by using Fourier analysis or by Green's theorem.

[1] In general dimension, explicit formulas are still available, but involve *Bessel functions*. But asymptotically at least, and ignoring higher order terms, one simply replaces $\frac{e^{ik|x-y|}}{4\pi|x-y|}$ by $\frac{e^{ik|x-y|}}{c_d|x-y|^{d-2}}$ for some explicit constant c_d.

Henceforth we restrict attention to three dimensions $d = 3$ for simplicity. One consequence of the above explicit formula is that for positive real $\lambda > 0$, the resolvents $R(\lambda + i\varepsilon)$ and $R(\lambda - i\varepsilon)$ tend to different limits as $\varepsilon \to 0$, reflecting the jump discontinuity in the resolvent function at the spectrum; as one can guess from formulae such as (6.4) or (6.5), such limits are of interest for understanding many other spectral multipliers. Indeed, for any test function f, we see that

$$\lim_{\varepsilon \to 0^+} R(\lambda + i\varepsilon) f(x) = \int_{\mathbf{R}^3} \frac{e^{i\sqrt{\lambda}|x-y|}}{4\pi|x-y|} f(y)\ dy$$

and

$$\lim_{\varepsilon \to 0^+} R(\lambda - i\varepsilon) f(x) = \int_{\mathbf{R}^3} \frac{e^{-i\sqrt{\lambda}|x-y|}}{4\pi|x-y|} f(y)\ dy.$$

Both of these functions

$$u_\pm(x) := \int_{\mathbf{R}^3} \frac{e^{\pm i\sqrt{\lambda}|x-y|}}{4\pi|x-y|} f(y)\ dy$$

solve the *Helmholtz equation*

$$(-\Delta - \lambda) u_\pm = f, \tag{6.6}$$

but have different asymptotics at infinity. Indeed, if $\int_{\mathbf{R}^3} f(y)\ dy = A$, then we have the asymptotic

$$u_\pm(x) = \frac{A e^{\pm i\sqrt{\lambda}|x|}}{4\pi|x|} + O(\frac{1}{|x|^2}) \tag{6.7}$$

as $|x| \to \infty$, leading also to the *Sommerfeld radiation condition*

$$u_\pm(x) = O(\frac{1}{|x|}); \quad (\partial_r \mp i\sqrt{\lambda}) u_\pm(x) = O(\frac{1}{|x|^2}) \tag{6.8}$$

where $\partial_r := \frac{x}{|x|} \cdot \nabla_x$ is the outgoing radial derivative. Indeed, one can show using an integration by parts argument that $u_\pm$ is the unique solution of the Helmholtz equation (6.6) obeying (6.8) (see below). u_+ is known as the *outward radiating* solution of the Helmholtz equation (6.6), and u_- is known as the *inward radiating* solution. Indeed, if one views the function $u_\pm(t,x) := e^{-i\lambda t} u_\pm(x)$ as a solution to the inhomogeneous Schrödinger equation

$$(i\partial_t + \Delta) u_\pm = -e^{-i\lambda t} f$$

and using the de Broglie law that a solution to such an equation with wave number $k \in \mathbf{R}^3$ (i.e. resembling $Ae^{ik\cdot x}$ for some amplitide A) should propagate at (group) velocity $2k$, we see (heuristically, at least) that the outward radiating solution will indeed propagate radially away from the origin at speed $2\sqrt{\lambda}$, while inward radiating solution propagates inward at the same speed.

There is a useful quantitative version of the convergence

$$R(\lambda \pm i\varepsilon)f \to u_\pm, \tag{6.9}$$

known as the *limiting absorption principle*:

Theorem 6.1.2 (Limiting absorption principle)**.** *Let f be a test function on $\mathbf{R}^3$, let $\lambda > 0$, and let $\sigma > 0$. Then one has*

$$\|R(\lambda \pm i\varepsilon)f\|_{H^{0,-1/2-\sigma}(\mathbf{R}^3)} \leq C_\sigma \lambda^{-1/2} \|f\|_{H^{0,1/2+\sigma}(\mathbf{R}^3)}$$

for all $\varepsilon > 0$, where $C_\sigma > 0$ depends only on σ, and $H^{0,s}(\mathbf{R}^3)$ is the weighted norm

$$\|f\|_{H^{0,s}(\mathbf{R}^3)} := \|\langle x\rangle^s f\|_{L^2_x(\mathbf{R}^3)}$$

and $\langle x\rangle := (1+|x|^2)^{1/2}$.

This principle allows one to extend the convergence (6.9) from test functions f to all functions in the weighted space $H^{0,1/2+\sigma}$ by a density argument (though the radiation condition (6.8) has to be adapted suitably for this scale of spaces when doing so). The weighted space $H^{0,-1/2-\sigma}$ on the left-hand side is optimal, as can be seen from the asymptotic (6.7); a duality argument similarly shows that the weighted space $H^{0,1/2+\sigma}$ on the right-hand side is also optimal.

We will prove this theorem shortly. As observed long ago by Kato [**Ka1965**] (and also reproduced below), this estimate is equivalent (via a Fourier transform in the spectral variable λ) to a useful estimate for the free Schrödinger equation known as the *local smoothing estimate*, which in particular implies the well-known *RAGE theorem* for that equation; it also has similar consequences for the free wave equation. As we shall see, it also encodes some spectral information about the Laplacian; for instance, it can be used to show that the Laplacian has no eigenvalues, resonances, or singular continuous spectrum. These spectral facts are already obvious from the Fourier transform representation of the Laplacian, but the point is that the limiting absorption principle also applies to more general operators for which the explicit diagonalisation afforded by the Fourier transform is not available; see [**RoTa2011**].

Important caveat: In order to illustrate the main ideas and suppress technical details, I will be a little loose with some of the rigorous details of the arguments, and in particular will be manipulating limits and integrals at a somewhat formal level.

6.1.1. Uniqueness. We first use an integration by parts argument to show uniqueness of the solution to the Helmholtz equation (6.6) assuming the radiation condition (6.8). For sake of concreteness we shall work with the sign $\pm = +$, and we will ignore issues of regularity, assuming all functions

are as smooth as needed. (In practice, the elliptic nature of the Laplacian ensures that issues of regularity are easily dealt with.) If uniqueness fails, then by subtracting the two solutions, we obtain a non-trivial solution u to the homogeneous Helmholtz equation

$$(-\Delta - \lambda)u = 0 \tag{6.10}$$

such that

$$u(x) = O(\frac{1}{|x|}); \quad (\partial_r - i\sqrt{\lambda})u(x) = O(\frac{1}{|x|^2}).$$

Next, we introduce the *charge current*

$$\mathbf{j}^i := \mathrm{Im}(\overline{u}\partial^i u)$$

(using the usual Einstein index notations), and observe from (6.6) that this current is divergence-free:

$$\partial_i \mathbf{j}^i = 0.$$

(This reflects the phase rotation invariance $u \mapsto e^{i\theta}u$ of the equation (6.6), and can also be viewed as a version of the conservation of the *Wronskian*.) From Stokes' theorem, and using polar coordinates, we conclude in particular that

$$\int_{S^2} \mathbf{j}^r(r\omega)\ d\omega = 0$$

or in other words that

$$\int_{S^2} \mathrm{Im}(\overline{u}\partial_r u)(r\omega)\ d\omega = 0.$$

Using the radiation condition, this implies in particular that

$$\int_{S^2} |u(r\omega)|^2\ d\omega = O(r^{-3}) \tag{6.11}$$

and

$$\int_{S^2} |\partial_r u(r\omega)|^2\ d\omega = O(r^{-3}) \tag{6.12}$$

as $r \to \infty$.

Now we use the "positive commutator method". Consider the expression

$$\int_{\mathbf{R}^3} [\partial_r, -\Delta - \lambda]u(x)\overline{u(x)}\ dx. \tag{6.13}$$

(To be completely rigorous, one should insert a cutoff to a large ball, and then send the radius of that ball to infinity, in order to make the integral well-defined but we will ignore this technicality here.) On the one hand, we may integrate by parts (using (6.11), (6.12) to show that all boundary terms

go to zero) and (6.10) to see that this expression vanishes. On the other hand, by expanding the Laplacian in polar coordinates we see that

$$[-\Delta - \lambda, \partial_r] = -\frac{2}{r^2}\partial_r - \frac{2}{r^3}\Delta_\omega.$$

An integration by parts in polar coordinates (using (6.11), (6.12) to justify ignoring the boundary terms at infinity) shows that

$$-\int_{\mathbf{R}^3} \frac{2}{r^2}\partial_r u(x)\overline{u(x)}\ dx = 8\pi|u(0)|^2$$

and

$$-\int_{\mathbf{R}^3} \frac{2}{r^3}\Delta_\omega u(x)\overline{u(x)}\ dx = 2\int_{\mathbf{R}^3} \frac{|\nabla_{\mathrm{ang}}|u(x)|^2}{|x|}\ dx$$

where $|\nabla_{\mathrm{ang}} u(x)|^2 := |\nabla u(x)|^2 - |\partial_r u(x)|^2$ is the angular part of the kinetic energy density $|\nabla u(x)|^2$. We obtain (a degenerate case of) the *Pohazaev-Morawetz identity*

$$8\pi|u(0)|^2 + 2\int_{\mathbf{R}^3} \frac{|\nabla_{\mathrm{ang}} u(x)|^2}{|x|}\ dx = 0$$

which implies in particular that u vanishes at the origin. Translating u around (noting that this does not affect either the Helmholtz equation or the Sommerfeld radiation condition) we see that u vanishes completely. (Alternatively, one can replace ∂_r by the smoothed out multiplier $\frac{x\cdot\nabla}{\langle x\rangle}$, in which case the Pohazaev-Morawetz identity acquires a term of the form $\int_{\mathbf{R}^3} \frac{|u(x)|^2}{\langle x\rangle^5}\ dx$ which is enough to directly ensure that u vanishes.)

6.1.2. Proof of the limiting absorption principle. We now sketch a proof of the limiting absorption principle, also based on the positive commutator method. For notational simplicity we shall only consider the case when λ is comparable to 1, though the method we give here also yields the general case after some more bookkeeping.

Let $\sigma > 0$ be a small exponent to be chosen later, and let f be normalised to have $H^{0,1/2+\sigma}(\mathbf{R}^3)$ norm equal to 1. For sake of concreteness let us take the + sign, so that we wish to bound $u := R(\lambda + i\varepsilon)f$. This u obeys the Helmholtz equation

$$\Delta u + \lambda u = f - i\varepsilon u. \tag{6.14}$$

For positive ε, we also see from the spectral theorem that u lies in $L^2(\mathbf{R}^3)$; the bound here though depends on ε, so we can only use this $L^2(\mathbf{R}^3)$ regularity for qualitative purposes (and specifically, for ensuring that boundary terms at infinity from integration by parts vanish) rather than quantitatively.

Once again, we may apply the positive commutator method. If we again consider the expression (6.13), then on the one hand this expression evaluates as before to

$$8\pi|u(0)|^2 + 2\int_{\mathbf{R}^3} \frac{|\nabla_{\mathrm{ang}} u(x)|^2}{|x|}\, dx.$$

On the other hand, integrating by parts using (6.14), this expression also evaluates to

$$2\,\mathrm{Re}\int_{\mathbf{R}^3} (\partial_r(-f + i\varepsilon u))\overline{u}\, dx.$$

Integrating by parts and using Cauchy-Schwarz and the normalisation on f (and also *Hardy's inequality*), we thus see that

$$|u(0)|^2 + \int_{\mathbf{R}^3} \frac{|\nabla_{\mathrm{ang}} u(x)|^2}{|x|}\, dx \lesssim \|u\|_{H^{0,-3/2-\sigma}} + \|\partial_r u\|_{H^{0,-1/2-\sigma}} + \varepsilon\|u\|_{L^2}\|\nabla u\|_{L^2}.$$

A slight modification of this argument, replacing the operator ∂_r with the smoothed out variant

$$(\frac{r}{\langle r\rangle} - \sigma\frac{r}{\langle r\rangle^{1+2\sigma}})\partial_r$$

yields (after a tedious computation)

$$\int_{\mathbf{R}^3} \frac{|u(x)|^2}{\langle x\rangle^{3+2\sigma}} + \frac{|\nabla u(x)|^2}{\langle x\rangle^{1+2\sigma}}\, dx \lesssim \|u\|_{H^{0,-3/2-\sigma}} + \|\partial_r u\|_{H^{0,-1/2-\sigma}} + \varepsilon\|u\|_{L^2}^2.$$

The left-hand side is $\|u\|_{H^{0,-3/2-\sigma}}^2 + \|\nabla u\|_{H^{0,-1/2-\sigma}}^2$; we can thus absorb the first two terms of the right-hand side onto the left-hand side, leading one with

$$\|u\|_{H^{0,-3/2-\sigma}}^2 + \|\nabla u\|_{H^{0,-1/2-\sigma}}^2 \lesssim \varepsilon\|u\|_{L^2}\|\nabla u\|_{L^2}.$$

On the other hand, by taking the inner product of (6.14) against iu and using the self-adjointness of $\Delta + \lambda$, one has

$$0 = \int_{\mathbf{R}^3} f\overline{iu} - \varepsilon\int_{\mathbf{R}^3} |u|^2$$

and hence by Cauchy-Schwarz and the normalisation of f

$$\varepsilon\|u\|_{L^2}^2 \leq \|u\|_{H^{0,-1/2-\sigma}}.$$

Elliptic regularity estimates using (6.14) (together with the hypothesis that λ is comparable to 1) also show that

$$\|u\|_{H^{0,-1/2-\sigma}} \lesssim \|\nabla u\|_{H^{0,-1/2-\sigma}} + 1$$

and

$$\|\nabla u\|_{L^2} \lesssim \|u\|_{L^2} + 1;$$

putting all these estimates together, we obtain

$$\|u\|_{H^{0,-1/2-\sigma}} \lesssim 1$$

as required.

Remark 6.1.3. In applications it is worth noting some additional estimates that can be obtained by variants of the above method (i.e. lots of integration by parts and Cauchy-Schwarz). From the Pohazaev-Morawetz identity, for instance, we can show some additional decay for the angular derivative:

$$\||\nabla_{\mathrm{ang}} u|\|_{H^{0,-1/2}} \lesssim \|f\|_{H^{0,1/2+\sigma}}.$$

With the positive sign $\pm = +$, we also have the Sommerfeld type outward radiation condition

$$\|\partial_r u - i\sqrt{\lambda} u\|_{H^{0,-1/2+\sigma}} \lesssim \|f\|_{H^{0,1/2+\sigma}}$$

if $\sigma > 0$ is small enough. For the negative sign $\pm = -$, we have the inward radiating condition

$$\|\partial_r u + i\sqrt{\lambda} u\|_{H^{0,-1/2+\sigma}} \lesssim \|f\|_{H^{0,1/2+\sigma}}$$

6.1.3. Spectral applications. The limiting absorption principle can be used to deduce various basic facts about the spectrum of the Laplacian. For instance:

Proposition 6.1.4 (Purely absolutely continuous spectrum)**.** *As an operator on $L^2(\mathbf{R}^3)$, $-\Delta$ has only purely absolutely continuous spectrum on any compact subinterval $[a,b]$ of $(0,+\infty)$.*

Proof. (Sketch) By density, it suffices to show that for any test function $f \in C_0^\infty(\mathbf{R}^3)$, the spectral measure μ_f of $-\Delta$ relative to f is purely absolutely continuous on $[a,b]$. In view of (6.5), we have

$$\mu_f = \lim_{\varepsilon \to 0^+} \frac{1}{\pi} \mathrm{Im} \langle R(\cdot + i\varepsilon) f, f \rangle$$

in the sense of distributions, so from Fatou's lemma it suffices to show that $\mathrm{Im}\langle R(\cdot + i\varepsilon) f, f\rangle$ is uniformly bounded on $[a,b]$, uniformly in ε. But this follows from the limiting absorption principle and Cauchy-Schwarz. □

Remark 6.1.5. The Laplacian $-\Delta$ also has no (point) spectrum at zero or negative energies, but this cannot be shown purely from the limiting absorption principle; if one allows a non-zero potential, then the limiting absorption principle holds (assuming suitable "short-range" hypotheses on the potential) but (as is well known in quantum mechanics) one can have eigenvalues (bound states) at zero or negative energies.

6.1.4. Local smoothing. Another key application of the limiting absorption principle is to obtain *local smoothing estimates* for both the Schrödinger and wave equations. Here is an instance of local smoothing for the Schrödinger equation:

Theorem 6.1.6 (Homogeneous local smoothing for Schrödinger)**.** *If $f \in L^2(\mathbf{R}^3)$, and $u : \mathbf{R} \times \mathbf{R}^3 \to \mathbf{C}$ is the (tempered distributional) solution to the homogeneous Schrödinger equation $iu_t + \Delta u = 0$, $u(0) = f$ (or equivalently, $u(t) = e^{it\Delta} f$), then one has*

$$\| |\nabla|^{1/2} u \|_{L^2_t H^{0,-1/2-\sigma}_x(\mathbf{R}\times\mathbf{R}^3)} \lesssim \|f\|_{L^2(\mathbf{R}^3)}$$

for any fixed $\sigma > 0$.

The $|\nabla|^{1/2}$ factor in this estimate is the "smoothing" part of the local smoothing estimate, while the negative weight $-1/2-\sigma$ is the "local" part. There is also a version of this local smoothing estimate for the inhomogeneous Schrödinger equation $iu_t + \Delta u = F$ which is in fact essentially *equivalent* to the limiting absorption principle (as observed *http://www.ams.org/mathscinet-getitem?mr=190801* by Kato), which we will not give here.

Proof. We begin by using the TT^* method. By duality, the claim is equivalent to

$$\| |\nabla|^{1/2} \int_{\mathbf{R}} e^{-it\Delta} F(t)\ dt \|_{L^2(\mathbf{R}^3)} \lesssim \|F\|_{L^2_t H^{0,1/2+\sigma}_x(\mathbf{R}\times\mathbf{R}^3)}$$

which by squaring is equivalent to

$$\| |\nabla| \int_{\mathbf{R}} e^{i(t-t')\Delta} F(t')\ dt' \|_{L^2_t H^{0,-1/2-\sigma}_x(\mathbf{R}\times\mathbf{R}^3)} \lesssim \|F\|_{L^2_t H^{0,1/2+\sigma}_x(\mathbf{R}\times\mathbf{R}^3)}. \tag{6.15}$$

From (6.5) one has (formally, at least)

$$e^{it\Delta} = \lim_{\varepsilon\to 0^+} \frac{1}{\pi} \int_{\mathbf{R}} (\operatorname{Im} R(y+i\varepsilon)) e^{-ity}\ dy.$$

Because $-\Delta$ only has spectrum on the positive real axis, $\operatorname{Im} R(y+i0)$ vanishes on the negative real axis, and so (after carefully dealing with the contribution near the zero energy) one has

$$e^{it\Delta} = \lim_{\varepsilon\to 0^+} \frac{1}{\pi} \int_0^\infty (\operatorname{Im} R(y+i\varepsilon)) e^{-ity}\ dy.$$

Taking the time-Fourier transform

$$\hat{F}(y) := \int_{\mathbf{R}} e^{ity} F(t)$$

we thus have

$$\int_{\mathbf{R}} e^{i(t-t')\Delta} F(t')\ dt' = \lim_{\varepsilon\to 0^+} \frac{1}{\pi} \int_{\mathbf{R}} e^{-ity} (\operatorname{Im} R(y+i\varepsilon)) \hat{F}(y)\ dy.$$

Applying Plancherel's theorem and Fatou's lemma (and commuting the L^2_t and $H^{0,-1/2-\sigma}_x$ norms), we can bound the LHS of (6.15) by

$$\lesssim \| |\nabla| (\operatorname{Im} R(y+i\varepsilon)) \hat{F}(y) \|_{L^2_y H^{0,-1/2-\sigma}_x(\mathbf{R}\times\mathbf{R}^3)}$$

while the right-hand side is comparable to

$$\lesssim \|\hat{F}(y)\|_{L^2_y H^{0,1/2+\sigma}_x(\mathbf{R}\times\mathbf{R}^3)}.$$

The claim now follows from the limiting absorption principle (and elliptic regularity). □

Remark 6.1.7. The above estimate was proven by taking a Fourier transform in time, and then applying the limiting absorption principle, which was in turn proven by using the positive commutator method. An equivalent way to proceed is to establish the local smoothing estimate directly by the analogue of the positive commutator method for Schrödinger flows, namely *Morawetz multiplier method* in which one contracts the stress-energy tensor (or variants thereof) against well-chosen vector fields, and integrates by parts.

An analogous claim holds for solutions to the wave equation

$$-\partial_t^2 u + \Delta u = 0$$

with initial data $u(0) = u_0$, $\partial_t u(0) = u_1$, with the relevant estimate being that

$$\||\nabla_{t,x} u|\|_{L^2_t H^{0,-1/2-\sigma}(\mathbf{R}\times\mathbf{R}^3)} \lesssim \|u_0\|_{H^1(\mathbf{R}^3)} + \|u_1\|_{L^2(\mathbf{R}^3)}.$$

As before, this estimate can also be proven directly using the Morawetz multiplier method.

6.1.5. The RAGE theorem. Another consequence of limiting absorption, closely related both to absolutely continuous spectrum and to local smoothing, is the RAGE theorem (named after Ruelle [**Ru1969**], Amrein-Georgescu [**AmGe1973**], and Enss [**En1978**], specialised to the free Schrödinger equation:

Theorem 6.1.8 (RAGE for Schrödinger)**.** *If $f \in L^2(\mathbf{R}^3)$, and K is a compact subset of $\mathbf{R}^3$, then $\|e^{it\Delta} f\|_{L^2(K)} \to 0$ as $t \to \pm\infty$.*

Proof. By a density argument we may assume that f lies in, say, $H^2(\mathbf{R}^3)$. Then $e^{it\Delta} f$ is uniformly bounded in $H^2(\mathbf{R}^3)$, and is Lipschitz in time in the $L^2(\mathbf{R}^3)$ (and hence $L^2(K)$) norm. On the other hand, from local smoothing we know that $\int_T^{T+1} \|e^{it\Delta}\|_{L^2(K)}\, dt$ goes to zero as $T \to \pm\infty$. Putting the two facts together we obtain the claim. □

Remark 6.1.9. One can also deduce this theorem from the fact that $-\Delta$ has purely absolutely continuous spectrum, using the abstract form of the RAGE theorem due to the authors listed above (which can be thought of as a Hilbert space-valued version of the *Riemann-Lebesgue lemma*).

There is also a similar RAGE theorem for the wave equation (with L^2 replaced by the energy space $H^1 \times L^2$) whose precise statement we omit here.

6.1.6. The limiting amplitude principle. A close cousin to the limiting absorption principle, which governs the limiting behaviour of the resolvent as it approaches the spectrum, is the *limiting amplitude principle*, which governs the asymptotic behaviour of a Schrödinger or wave equation with oscillating forcing term. We give this principle for the Schrödinger equation (the case for the wave equation is analogous):

Theorem 6.1.10 (Limiting amplitude principle)**.** *Let $f \in L^2(\mathbf{R}^3)$ be compactly supported, let $\mu > 0$, and let u be a solution to the forced Schrödinger equation $iu_t + \Delta u = e^{-i\mu t} f$ which lies in $L^2(\mathbf{R}^3)$ at time zero. Then for any compact set K, $e^{i\mu T} u$ converges in $L^2(K)$ as $T \to +\infty$ to v, the solution to the Helmholtz equation $\Delta v + \mu v = f$ obeying the outgoing radiation condition* (6.7).

Proof. (Sketch) By subtracting off the free solution $e^{it\Delta} u(0)$ (which decays in $L^2(K)$ by the RAGE theorem), we may assume that $u(0) = 0$. From the Duhamel formula we then have

$$u(T) = -i \int_0^T e^{i(T-t)\Delta} e^{-i\mu t} f \, dt$$

and thus (after changing variables from t to $T - t$)

$$e^{i\mu T} u(T) = -i \int_0^T e^{it(\Delta+\mu)} f \, dt.$$

We write the right-hand side as

$$-i \lim_{\varepsilon \to 0^+} \int_0^T e^{it(\Delta+\mu+i\varepsilon)} f \, dt.$$

From the limiting absorption principle, the integral $-i \int_0^\infty e^{it(\Delta+\mu+i\varepsilon)} f \, dt$ converges to v, and so it suffices to show that the expression

$$\lim_{\varepsilon \to 0^+} \int_T^\infty e^{it(\Delta+\mu+i\varepsilon)} f \, dt$$

converges to zero as $T \to +\infty$ in $L^2(K)$ norm. Evaluating the integral, we are left with showing that

$$\lim_{\varepsilon \to 0^+} e^{iT\Delta} R(\mu + i\varepsilon) f$$

converges to zero as $T \to +\infty$ in $L^2(K)$ norm.

By using contour integration, one can write

$$\lim_{\varepsilon\to 0^+} e^{iT\Delta} R(\mu+i\varepsilon)f = \frac{1}{2\pi i}\lim_{\varepsilon\to 0^+}\lim_{\varepsilon'\to 0^+}\int_{\mathbf{R}} \frac{e^{-iTx}}{x-\mu-i\varepsilon} R(x+i\varepsilon')f\ dx.$$

On the other hand, from the explicit solution for the resolvent (and the compact support of f), $R(x+i\varepsilon')f$ can be shown to vary in a Hölder continuous fashion on x in the $L^2(K)$ norm (uniformly in x, ε'), and to decay at a polynomial rate as $x \to \pm\infty$. Since

$$\int_{\mathbf{R}} \frac{e^{-iTx}}{x-\mu-i\varepsilon}\ dx = 0$$

for $T > 0$, the required decay in $L^2(K)$ then follows from a routine calculation. □

Remark 6.1.11. More abstractly, it was observed by Eidus [**Ei1969**] that the limiting amplitude principle for a general Schrödinger or wave equation can be deduced from the limiting absorption principle and a Hölder continuity bound on the resolvent.

6.2. The shallow water wave equation, and the propagation of tsunamis

Tsunamis are water waves that start in the deep ocean, usually because of an underwater earthquake (though tsunamis can also be caused by underwater landslides or volcanoes), and then propagate towards shore. Initially, tsunamis have relatively small amplitude (a metre or so is typical), which would seem to render them as harmless as *wind waves*. And indeed, tsunamis often pass by ships in deep ocean without anyone on board even noticing.

However, being generated by an event as large as an earthquake, the *wavelength* of the tsunami is huge - 200 kilometres is typical (in contrast with wind waves, whose wavelengths are typically closer to 100 metres). In particular, the wavelength of the tsunami is far greater than the depth of the ocean (which is typically 2-3 kilometres). As such, even in the deep ocean, the dynamics of tsunamis are essentially governed by the *shallow water equations*. One consequence of these equations is that the speed of propagation v of a tsunami can be approximated by the formula

$$v \approx \sqrt{gb} \tag{6.16}$$

where b is the depth of the ocean, and $g \approx 9.8 ms^{-2}$ is the force of gravity. As such, tsunamis in deep water move[2] *very* fast - speeds such as 500 kilometres per hour (300 miles per hour) are quite typical; enough to travel from Japan to the US, for instance, in less than a day. Ultimately, this is

[2]Note though that this is the *phase velocity* of the tsunami wave, and not the velocity of the water molecues themselves, which are far slower.

due to the incompressibility of water (and conservation of mass); the massive net pressure (or more precisely, spatial variations in this pressure) of a very broad and deep wave of water forces the profile of the wave to move horizontally at vast speeds.

As the tsunami approaches shore, the depth b of course decreases, causing the tsunami to slow down, at a rate proportional to the square root of the depth, as per (6.16). Unfortunately, *wave shoaling* then forces the amplitude A to increase at an inverse rate governed by *Green's law*,

$$A \propto \frac{1}{b^{1/4}} \tag{6.17}$$

at least until the amplitude becomes comparable to the water depth (at which point the assumptions that underlie the above approximate results break down; also, in two (horizontal) spatial dimensions there will be some decay of amplitude as the tsunami spreads outwards). If one starts with a tsunami whose initial amplitude was A_0 at depth b_0 and computes the point at which the amplitude A and depth b become comparable using the proportionality relationship (6.17), some high school algebra then reveals that at this point, amplitude of a tsunami (and the depth of the water) is about $A_0^{4/5} b_0^{1/5}$. Thus, for instance, a tsunami with initial amplitude of one metre at a depth of 2 kilometres can end up with a final amplitude of about 5 metres near shore, while still traveling at about ten metres per second (35 kilometres per hour, or 22 miles per hour), which can lead to a devastating impact when it hits shore.

While tsunamis are far too massive of an event to be able to control (at least in the deep ocean), we can at least model them mathematically, allowing one to predict their impact at various places along the coast with high accuracy. The full equations and numerical methods used to perform such models are somewhat sophisticated, but by making a large number of simplifying assumptions, it is relatively easy to come up with a rough model that already predicts the basic features of tsunami propagation, such as the velocity formula (6.16) and the amplitude proportionality law (6.17). I give this (standard) derivation below. The argument will largely be heuristic in nature; there are very interesting analytic issues in actually justifying many of the steps below rigorously, but I will not discuss these matters here.

6.2.1. The shallow water wave equation. The ocean is, of course, a three-dimensional fluid, but to simplify the analysis we will consider a two-dimensional model in which the only spatial variables are the horizontal variable x and the vertical variable z, with $z = 0$ being equilibrium sea level. We model the ocean floor by a curve

$$z = -b(x),$$

thus b measures the depth of the ocean at position x. At any time t and position x, the height of the water (compared to sea level $z = 0$) will be given by an unknown height function $h(t, x)$; thus, at any time t, the ocean occupies the region

$$\Omega_t := \{(x, z) : -b(x) < z < h(t, x)\}.$$

Now we model the motion of water inside the ocean by assigning at each time t and each point $(x, z) \in \Omega_t$ in the ocean, a velocity vector

$$\vec{u}(t, x, z) = (u_x(t, x, z), u_z(t, x, z)).$$

We make the basic assumption of *incompressibility*, so that the density ρ of water is constant throughout Ω_t.

The velocity changes over time according to Newton's second law $F = ma$. To apply this law to fluids, we consider an infinitesimal amount of water as it flows along the velocity field $\vec{u}$. Thus, at time t, we assume that this amount of water occupies some infinitesimal area dA and some position $\vec{x}(t) = (x(t), z(t))$, where we have

$$\frac{d}{dt}\vec{x}(t) = \vec{u}(t, \vec{x}(t)).$$

Because of incompressibility, the area dA stays constant, and the mass of this infinitesimal portion of water is $m = \rho dA$. There will be two forces on this body of water; the force of gravity, which is $(0, -mg) = (0, -\rho)dA$, and the force of the pressure field $p(t, x, z)$, which is given by $-\nabla p dA$. At the length and time scales of a tsunami, we can neglect the effect of other forces such as viscosity or surface tension. Newton's law $m\frac{du}{dt} = F$ then gives

$$m\frac{d}{dt}\vec{u}(t, \vec{x}(t)) = -\nabla p dA + (0, -mg)$$

which simplifies to the *incompressible Euler equation*

$$\frac{\partial}{\partial t}\vec{u} + (\vec{u} \cdot \nabla)\vec{u} = -\frac{1}{\rho}\nabla p + (0, -g).$$

At present, the pressure is not given. However, we can simplify things by making the assumption of *(vertical) hydrostatic equilibrium*, i.e. the vertical effect $-\frac{1}{\rho}\frac{\partial}{\partial z}p$ of pressure cancels out the effect $-g$ of gravity. We also assume that the pressure is zero on the surface $z = h(t, x)$ of the water. Together, these two assumptions force the pressure to be the *hydrostatic pressure*

$$p = \rho g(h(t, x) - z). \tag{6.18}$$

This reflects the intuitively plausible fact that the pressure at a point under the ocean should be determined by the weight of the water above that point.

The incompressible Euler equation now simplifies to

$$\frac{\partial}{\partial t}\vec{u} + (\vec{u}\cdot\nabla)\vec{u} = -g(\frac{\partial}{\partial x}h, 0). \tag{6.19}$$

We next make the *shallow water* approximation that the wavelength of the water is far greater than the depth of the water. In particular, we do not expect significant changes in the velocity field in the z variable, and thus make the ansatz

$$\vec{u}(t,x,z) \approx \vec{u}(t,x). \tag{6.20}$$

(This ansatz should be taken with a grain of salt, particularly when applied to the z component u_z of the velocity, which does actually have to fluctuate a little bit to accomodate changes in ocean depth and in the height function. However, the primary component of the velocity is the horizontal component u_x, and this does behave in a fairly vertically insensitive fashion in actual tsunamis.)

Taking the x component of (6.19), and abbreviating u_x as u, we obtain the first shallow water wave equation

$$\frac{\partial}{\partial t}u + u\frac{\partial}{\partial x}u = -g\frac{\partial}{\partial x}h. \tag{6.21}$$

The next step is to play off the incompressibility of water against the finite depth of the ocean. Consider an infinitesimal slice

$$\{(x,z) \in \Omega_t : x_0 \leq x \leq x_0 + dx\}$$

of the ocean at some time t and position x_0. The total mass of this slice is roughly

$$\rho(h(t,x_0) + b(x_0))dx$$

and so the rate of change of mass of this slice over time is

$$\rho\frac{\partial h}{\partial t}(t,x_0)dx.$$

On the other hand, the rate of mass entering this slice on the left $x = x_0$ is

$$\rho u(t,x_0)(h(t,x_0) + b(x_0))$$

and the rate of mass exiting on the right $x = x_0 + dx$ is

$$\rho u(t,x_0+dx)(h(t,x_0+dx) + b(x_0+dx)).$$

Putting these three facts together, we obtain the equation

$$\begin{aligned}\rho\frac{\partial h}{\partial t}(t,x_0)dx = {}& \rho u(t,x_0)(h(t,x_0) + b(x_0)) \\ & - \rho u(t,x_0+dx)(h(t,x_0+dx) + b(x_0+dx))\end{aligned}$$

which simplifies after Taylor expansion to the second shallow water wave equation

$$\frac{\partial}{\partial t}h + \frac{\partial}{\partial x}(u(h+b)) = 0. \tag{6.22}$$

Remark 6.2.1. Another way to derive (6.22) is to use a more familiar form of the incompressibility, namely the divergence-free equation

$$\frac{\partial}{\partial x}u_x + \frac{\partial}{\partial z}u_z = 0. \tag{6.23}$$

(Here we will refrain from applying (6.20) to the vertical component of the velocity u_z, as the approximation (6.20) is not particularly accurate for this component.) Also, by considering the trajectory of a particle $(x(t), h(t, x(t)))$ at the surface of the ocean, we have the formulae

$$\frac{d}{dt}x(t) = u_x(x(t), h(t, x(t)))$$

and

$$\frac{d}{dt}h(t, x(t)) = u_z(x(t), h(t, x(t)))$$

which after application of the chain rule gives the equation

$$\frac{\partial}{\partial t}h(t,x) + (\frac{\partial}{\partial x}h(x))u_x(x, h(t,x)) = u_z(x, h(t,x)). \tag{6.24}$$

A similar analysis at the ocean floor (which does not vary in time) gives

$$-\frac{\partial}{\partial x}b(x)u_x(x, -b(x)) = u_z(x, -b(x)). \tag{6.25}$$

We apply these equations to the evaluation of the expression

$$\frac{\partial}{\partial x}\int_{-b(x)}^{h(t,x)} u_x(t,x,z)\ dz.$$

which is the spatial rate of change of the velocity flux through a vertical slice of the ocean. On the one hand, using the ansatz (6.20), we expect this expression to be approximately

$$\frac{\partial}{\partial x}(u(h+b)).$$

On the other hand, by differentiation under the integral sign, we can evaluate this expression instead as

$$\begin{aligned}\int_{-b(x)}^{h(t,x)} &\frac{\partial}{\partial x}u_x(t,x,z)\ dz \\ &+ (\frac{\partial}{\partial x}h(t,x))u_x(x,h(t,x)) \\ &+ (\frac{\partial}{\partial x}b(x))u_x(x,-b(x)).\end{aligned}$$

If we then substitute in (6.23), (6.24), (6.25) and apply the fundamental theorem of calculus, one ends up with $-\frac{\partial}{\partial t}h(t,x)$, and the claim (6.22) follows.

The equations (6.21), (6.22) are nonlinear in the unknowns u, h. However, one can approximately linearise them by making the hypothesis that the amplitude of the wave is small compared to the depth of the water:

$$|h| \ll b. \tag{6.26}$$

This hypothesis is fairly accurate for tsunamis in the deep ocean, and even for medium depths, but of course is not reasonable once the tsunami has reached shore (where the dynamics are far more difficult to model).

The hypothesis (6.26) already simplifies (6.22) to (approximately)

$$\frac{\partial}{\partial t}h + \frac{\partial}{\partial x}(ub) = 0. \tag{6.27}$$

As for (6.21), we argue that the second term on the left-hand side is negligible, leading to

$$\frac{\partial}{\partial t}u = -g\frac{\partial}{\partial x}h. \tag{6.28}$$

To explain heuristically why we expect this to be the case, let us make the ansatz that h and u have amplitude A, V respectively, and propagate at some *phase velocity* v and wavelength λ; let us also make the (reasonable) assumption that b varies much slower in space than u does (i.e. that b is roughly constant at the scale of the wavelength λ), so we may (for a first approximation) replace $\frac{\partial}{\partial x}(ub)$ by $b\frac{\partial}{\partial x}u$. Heuristically, we then have

$$\begin{aligned}
\frac{\partial}{\partial x}u &= O(V/\lambda) \\
\frac{\partial}{\partial x}h &= O(A/\lambda) \\
\frac{\partial}{\partial t}u &= O(vV/\lambda) \\
\frac{\partial}{\partial t}h &= O(vA/\lambda)
\end{aligned}$$

and equation (6.27) then suggests

$$vA/\lambda \approx Vb/\lambda. \tag{6.29}$$

From (6.26) we expect $A \ll b$, and thus $v \gg V$; the wave propagates much faster than the velocity of the fluid. In particular, we expect $u\frac{\partial}{\partial x}u = O(V^2/\lambda)$ to be much smaller than $\frac{\partial}{\partial t}u = O(vV/\lambda)$, which explains why we expect to drop the second term in (6.21) to obtain (6.28).

If we now insert the above ansatz into (6.28), we obtain

$$vV/\lambda \approx gA/\lambda;$$

combining this with (6.29), we already get the velocity relationship (6.16).

Remark 6.2.2. One can also obtain (6.16) more quickly (up to a loss of a constant factor) by *dimensional analysis*, together with some additional physical arguments. Indeed, it is clear from a superficial scan of the above discussion that the velocity v is only going to depend on the quantities $\rho, g, b, A, V, \lambda$. As the density ρ is the only input that involves mass in its units, dimensional analysis already rules out any role for ρ. As we are in the small amplitude regime (6.26), we expect the dynamics to be linearised, and thus not dependent on amplitude; this rules out A (and similarly V, which is the amplitude of the velocity field, and which is negligible when compared against the phase velocity V). Finally, in the long wavelength regime $\lambda \gg b$, we expect the wavelength to be physically undetectable at local scales (it requires not only knowledge of the slope of the height function at one's location, but also the second derivative of that function (i.e. the curvature of the ocean surface), which is lower order). So we rule out dependence on λ also, leaving only g and b, and at this point dimensional analysis forces the relationship (6.16) up to constants. (Unfortunately, I do not know of an analogous dimensional analysis argument that gives (6.17).)

To get the relation (6.17), we have to analyse the ansatz a bit more carefully. First, we combine (6.28) and (6.27) into a single equation for the height function h. Indeed, differentiating (6.27) in time and then substituting in (6.28) and (6.16) gives

$$\frac{\partial^2}{\partial t^2} h - \frac{\partial}{\partial x}(v^2 \frac{\partial}{\partial x} h) = 0.$$

To solve this wave equation, we use a standard sinusoidal ansatz

$$h(t,x) = A(t,x)\sin(\phi(t,x)/\varepsilon)$$

where A, ϕ are slowly varying functions, and $\varepsilon > 0$ is a small parameter. Inserting this ansatz and extracting the top order terms in ε, we conclude the eikonal equation

$$\phi_t^2 - v^2 \phi_x^2 = 0$$

and the Hamilton-Jacobi equation

$$2A_t\phi_t + A\phi_{tt} - v^2(2A_x\phi_x + A\phi_{xx}) - 2vv_xA\phi_x = 0.$$

From the eikonal equation we see that ϕ propagates at speed v. Assuming rightward propagation, we thus have

$$\phi_t = -v\phi_x. \tag{6.30}$$

As for the Hamilton-Jacobi equation, we solve it using the method of characteristics. Multiplying the equation by A, we obtain

$$(A^2\phi_t)_t - v^2(A^2\phi_x)_x - 2vv_xA^2\phi_x = 0.$$

Inserting (6.30) and writing $F := A^2\phi_x$, one obtains

$$-vF_t - v^2F_x - 2vv_xF = 0$$

which simplifies to

$$(\partial_t + v\partial_x)(v^2F) = 0.$$

Thus we see that v^2F is constant along characteristics. On the other hand, differentiating (6.30) in x we see (after some rearranging) that

$$(\partial_t + v\partial_x)(v\phi_x) = 0$$

so $v\phi_x$ is also constant along characteristics. Dividing, we see that A^2v is constant along characteristics, leading to the proportionality relationship

$$A \propto \frac{1}{\sqrt{v}}$$

which gives (6.17).

Remark 6.2.3. It becomes difficult to retain the sinusoidal ansatz once the amplitude exceeds the depth, as it leads to the absurd conclusion that the troughs of the wave lie below the ocean floor. However, a remnant of this effect can actually be seen in real-life tsunamis, namely that if the tsunami starts with a trough rather than a crest, then the water at the shore draws back at first (sometimes for hundreds of metres), before the crest of the tsunami hits. As such, the sudden withdrawal of water of a shore is an important warning sign of an immediate tsunami.

Chapter 7

Number theory

7.1. Hilbert's seventh problem, and powers of 2 and 3

Hilbert's seventh problem asks to determine the transcendence of powers a^b of two algebraic numbers a, b. This problem was famously solved by Gelfond and Schneider [**Ge1934**], [**Sc1934**]:

Theorem 7.1.1 (Gelfond-Schneider theorem)**.** *Let a, b be algebraic numbers, with $a \neq 0, 1$ and b irrational. Then (any of the values of the possibly multi-valued expression) a^b is transcendental.*

For sake of simplifying the discussion, let us focus on just one specific consequence of this theorem:

Corollary 7.1.2. $\frac{\log 2}{\log 3}$ *is transcendental.*

Proof. If not, one could obtain a contradiction to the Gelfond-Schneider theorem by setting $a := 3$ and $b := \frac{\log 2}{\log 3}$. (Note that $\frac{\log 2}{\log 3}$ is clearly irrational, since $3^p \neq 2^q$ for any integers p, q with q positive.) □

In a series of papers [**Ba1966**], [**Ba1967**], [**Ba1967b**], Alan Baker established a major generalisation of the Gelfond-Schneider theorem known as *Baker's theorem*, as part of his work in transcendence theory that later earned him a Fields Medal. Among other things, this theorem provided explicit quantitative bounds on exactly *how* transcendental quantities such as $\frac{\log 2}{\log 3}$ were. In particular, it gave a strong bound on how *irrational* such quantities were (i.e. how easily they were approximable by rationals). Here, in particular, is one special case of Baker's theorem:

Proposition 7.1.3 (Special case of Baker's theorem)**.** *For any integers p, q with q positive, one has*

$$|\frac{\log 2}{\log 3} - \frac{p}{q}| \geq c\frac{1}{q^C}$$

for some absolute (and effectively computable) constants $c, C > 0$.

This theorem may be compared with (the easily proved) Liouville's theorem on diophantine approximation, which asserts that if α is an irrational algebraic number of degree d, then

$$|\alpha - \frac{p}{q}| \geq c\frac{1}{q^d}$$

for all p, q with q positive, and some effectively computable $c > 0$, and (the more significantly difficult) *Thue-Siegel-Roth theorem* [**Th1909, Si1921, Ro1955**], which under the same hypotheses gives the bound

$$|\alpha - \frac{p}{q}| \geq c_\varepsilon\frac{1}{q^{2+\varepsilon}}$$

for all $\varepsilon > 0$, all p, q with q positive and an *ineffective*[1] constant $c_\varepsilon > 0$. Finally, one should compare these results against *Dirichlet's theorem on Diophantine approximation*, which asserts that for any real number α one has

$$|\alpha - \frac{p}{q}| < \frac{1}{q^2}$$

for infinitely many p, q with q positive.

Proposition 7.1.3 easily implies the following separation property between powers of 2 and powers of 3:

Corollary 7.1.4 (Separation between powers of 2 and powers of 3)**.** *For any positive integers p, q one has*

$$|3^p - 2^q| \geq \frac{c}{q^C}3^p$$

for some effectively computable constants $c, C > 0$ (which may be slightly different from those in Proposition 7.1.3).

Indeed, this follows quickly from Proposition 7.1.3, the identity

$$3^p - 2^q = 3^p(1 - 3^{q(\frac{\log 2}{\log 3} - \frac{p}{q})}) \tag{7.1}$$

and some elementary estimates.

In particular, the gap between powers of three 3^p and powers of two 2^q grows exponentially in the exponents p, q. I do not know of any other way

[1]The reason the Thue-Siegel-Roth theorem is ineffective is because it relies heavily on the dueling conspiracies argument [**Ta2010b**, §1.12], i.e. playing off multiple "conspiracies" $\alpha \approx \frac{p}{q}$ against each other; the other results however only focus on one approximation at a time and thus avoid ineffectivity.

to establish this fact other than essentially going through some version of Baker's argument (which will be given below).

For comparison, by exploiting the trivial (yet fundamental) *integrality gap* - the obvious fact that if an integer n is non-zero, then its magnitude is at least 1 - we have the trivial bound

$$|3^p - 2^q| \geq 1$$

for all positive integers p, q (since, from the fundamental theorem of arithmetic, $3^p - 2^q$ cannot vanish). Putting this into (7.1) we obtain a very weak version of Proposition 7.1.3, that only gives exponential bounds instead of polynomial ones:

Proposition 7.1.5 (Trivial bound)**.** *For any integers p, q with q positive, one has*

$$|\frac{\log 2}{\log 3} - \frac{p}{q}| \geq c\frac{1}{2^q}$$

for some absolute (and effectively computable) constant $c > 0$.

The proof of Baker's theorem (or even of the simpler special case in Proposition 7.1.3) is largely elementary (except for some very basic complex analysis), but is quite intricate and lengthy, as a lot of careful book-keeping is necessary in order to get a bound as strong as that in Proposition 7.1.3. To illustrate the main ideas, I will prove a bound that is weaker than Proposition 7.1.3, but still significantly stronger than Proposition 7.1.5, and whose proof already captures many of the key ideas of Baker:

Proposition 7.1.6 (Weak special case of Baker's theorem)**.** *For any integers p, q with $q > 1$, one has*

$$|\frac{\log 2}{\log 3} - \frac{p}{q}| \geq \exp(-C \log^{C'} q)$$

for some absolute constants $C, C' > 0$.

Note that Proposition 7.1.3 is equivalent to the assertion that one can take $C' = 1$ (and C effective) in the above proposition.

The proof of Proposition 7.1.6 can be made effective (for instance, it is not too difficult to make the C' close to 2); however, in order to simplify the exposition (and in particular, to be able to use some nonstandard analysis terminology to reduce the epsilon management, cf. [**Ta2008**, §1.5]), I will establish Proposition 7.1.6 with ineffective constants C, C'.

Like many other results in transcendence theory, the proof of Baker's theorem (and Proposition 7.1.6) rely on what we would nowadays call the *polynomial method* - to play off upper and lower bounds on the complexity of polynomials that vanish (or nearly vanish) to high order on a specified

set of points. In the specific case of Proposition 7.1.6, the points in question are of the form

$$\Gamma_N := \{(2^n, 3^n) : n = 1, \ldots, N\} \subset \mathbf{R}^2$$

for some large integer N. On the one hand, the irrationality of $\frac{\log 2}{\log 3}$ ensures that the curve

$$\gamma := \{(2^t, 3^t) : t \in \mathbf{R}\}$$

is not algebraic, and so it is difficult for a polynomial P of controlled complexity[2] to vanish (or nearly vanish) to high order at all the points of Γ_N; the trivial bound in Proposition 7.1.5 allows one to make this statement more precise. On the other hand, if Proposition 7.1.6 failed, then $\frac{\log 2}{\log 3}$ is close to a rational, which by Taylor expansion makes γ close to an algebraic curve over the rationals (up to some rescaling by factors such as $\log 2$ and $\log 3$) at each point of Γ_N. This, together with a pigeonholing argument, allows one to find a polynomial P of reasonably controlled complexity to (nearly) vanish to high order at every point of Γ_N.

These observations, by themselves, are not sufficient to get beyond the trivial bound in Proposition 7.1.5. However, Baker's key insight was to exploit the integrality gap to bootstrap the (near) vanishing of P on a set Γ_N to imply near-vanishing of P on a larger set $\Gamma_{N'}$ with $N' > N$. The point is that if a polynomial P of controlled degree and size (nearly) vanishes to higher order on a lot of points on an analytic curve such as γ, then it will also be fairly small on many other points in γ as well. (To quantify this statement efficiently, it is convenient to use the tools of complex analysis, which are particularly well suited to understand zeroes (or small values) of polynomials.) But then, thanks to the integrality gap (and the controlled complexity of P), we can amplify "fairly small" to "very small".

Using this observation and an iteration argument, Baker was able to take a polynomial of controlled complexity P that nearly vanished to high order on a relatively small set Γ_{N_0}, and bootstrap that to show near-vanishing on a much larger set Γ_{N_k}. This bootstrap allows one to dramatically bridge the gap between the upper and lower bounds on the complexity of polynomials that nearly vanish to a specified order on a given Γ_N, and eventually leads to Proposition 7.1.6 (and, with much more care and effort, to Proposition 7.1.3).

Below the fold, I give the details of this argument. My treatment here is inspired by the exposé [**Se1969**], as well as the unpublished lecture notes [**So2010**].

[2]Here, "complexity" of a polynomial is an informal term referring both to the degree of the polynomial, and the height of the coefficients, which in our application will essentially be integers up to some normalisation factors.

7.1.1. Nonstandard formulation. The proof of Baker's theorem requires a lot of "epsilon management" in that one has to carefully choose a lot of parameters such as C and ε in order to make the argument work properly. This is particularly the case if one wants a good value of exponents in the final result, such as the quantity C' in Proposition 7.1.6. To simplify matters, we will abandon all attempts to get good values of constants anywhere, which allows one to retreat to the nonstandard analysis setting where the notation is much cleaner, and much (though not all) of the epsilon management is eliminated (cf. [**Ta2008**, §1.5]). This is a relatively mild use of nonstandard analysis, though, and it is not difficult to turn all the arguments below into standard effective arguments (but at the cost of explicitly tracking all the constants C). See for instance [**So2010**] for such an effective treatment.

We turn to the details. We will assume some basic familiarity with nonstandard analysis, as covered for instance in [**Ta2008**, §1.5] (but one should be able to follow this argument using only non-rigorous intuition of what terms such as "unbounded" or "infinitesimal" mean).

Let H be an unbounded (nonstandard) positive real number. Relative to this H, we can define various notions of size:

(1) A nonstandard number z is said to be *of polynomial size* if one has $|z| \leq CH^C$ for some standard $C > 0$.

(2) A nonstandard number z is said to be *of polylogarithmic size* if one has $|z| \leq C \log^C H$ for some standard $C > 0$.

(3) A nonstandard number z is said to be *of quasipolynomial size* if one has $|z| \leq \exp(C \log^C H)$ for some standard $C > 0$.

(4) A nonstandard number z is said to be *quasiexponentially small* if one has $|z| \leq \exp(-C \log^C H)$ for every standard $C > 0$.

(5) Given two nonstandard numbers X, Y with Y non-negative, we write $X \ll Y$ or $X = O(Y)$ if $|X| \leq CY$ for some standard $C > 0$. We write $X = o(Y)$ or $X \lll Y$ if we have $|X| \leq cY$ for all standard $c > 0$.

As a general rule of thumb, in our analysis all exponents will be of polylogarithmic size, all coefficients will be of quasipolynomial size, and all error terms will be quasiexponentially small.

In this nonstandard analysis setting, there is a clean calculus (analogous to the calculus of the asymptotic notations $O()$ and $o()$) to manipulate these sorts of quantities without having to explicitly track the constants C. For instance:

(1) The sum, product, or difference of two quantities of a given size (polynomial, polylogarithmic, quasipolynomial, or quasiexponentially small) remains of that given size (i.e. each size range forms a ring).

(2) If $X \ll Y$, and Y is of a given size, then X is also of that size.

(3) If X is of quasipolynomial size and Y is of polylogarithmic size, then X^Y is of quasipolynomial size, and (if Y is a natural number) $Y!$ is also of quasipolynomial size.

(4) If ε is quasiexponentially small, and X is of quasipolynomial size, then $X\varepsilon$ is also quasiexponentially small. (Thus, the quasiexponentially small numbers form an ideal in the ring of quasipolynomial numbers.)

(5) Any quantity of polylogarithmic size, is of polynomial size; and any quantity of polynomial size, is of quasipolynomial size.

We will refer to these sorts of facts as *asymptotic calculus*, and rely upon them heavily to simplify a lot of computations (particularly regarding error terms).

Proposition 7.1.6 is then equivalent to the following assertion:

Proposition 7.1.7 (Nonstandard weak special case of Baker)**.** *Let H be an unbounded nonstandard natural number, and let $\frac{p}{q}$ be a rational of height at most H (i.e. $|p|, |q| \leq H$). Then $\frac{\log 2}{\log 3} - \frac{p}{q}$ is not quasiexponentially small (relative to H, of course).*

Let us quickly see why Proposition 7.1.7 implies Proposition 7.1.6 (the converse is easy and is left to the reader). This is the usual "compactness and contradiction" argument. Suppose for contradiction that Proposition 7.1.6 failed. Carefully negating the quantifiers, we may then find a sequence $\frac{p_n}{q_n}$ of (standard) rationals with $q_n > 1$, such that

$$|\frac{\log 2}{\log 3} - \frac{p_n}{q_n}| \geq \exp(-n \log^n q_n)$$

for all natural numbers n. As $\frac{\log 2}{\log 3}$ is irrational, q_n must go to infinity. Taking the ultralimit $\frac{p}{q}$ of the $\frac{p_n}{q_n}$, and setting H to be (say) q, we contradict Proposition 7.1.7.

It remains to prove Proposition 7.1.7. We fix the unbounded nonstandard natural number H, and assume for contradiction that $\frac{\log 2}{\log 3}$ is quasiexponentially close to a nonstandard rational $\frac{p}{q}$ of height at most H. We will write $X \approx Y$ for the assertion that $X - Y$ is quasiexponentially small, thus

$$\frac{\log 2}{\log 3} \approx \frac{p}{q}. \tag{7.2}$$

The objective is to show that (7.2) leads to a contradiction.

7.1.2. The polynomial method. Now it is time to introduce the polynomial method. We will be working with the following class of polynomials:

Definition 7.1.8. A *good polynomial* is a nonstandard polynomial $P : {}^*\mathbf{C}^2 \to {}^*\mathbf{C}$ of the form

$$P(x,y) = \sum_{0 \leq a,b \leq D} c_{a,b} x^a y^b \tag{7.3}$$

of two nonstandard variables of some (nonstandard) degree at most D (in each variable), where D is a nonstandard natural number of polylogarithmic size, and whose coefficients $c_{a,b}$ are (nonstandard) integers of quasipolynomial size. (A technical point: we require the $c_{a,b}$ to depend in an internal fashion on the indices a, b, in order for the nonstandard summation here to be well-defined.) Define the *height* M of the polynomial to be the maximum magnitude of the coefficients in P; thus, by hypothesis, M is of quasipolynomial size.

We have a key definition:

Definition 7.1.9. Let N, J be two (nonstandard) positive numbers of polylogarithmic size. A good polynomial P is said to *nearly vanish to order* J on Γ_N if one has

$$\frac{d^j}{dz^j} P(2^z, 3^z)|_{z=n} \approx 0 \tag{7.4}$$

for all nonstandard natural numbers $0 \leq j \leq J$ and $1 \leq n \leq N$.

The derivatives in (7.4) can be easily computed. Indeed, if we expand out the good polynomial P out as (7.3), then the left-hand side of (7.4) is

$$\sum_{0 \leq a,b \leq D} c_{a,b} (a \log 2 + b \log 3)^j 2^{an} 3^{bn}.$$

Now, from (7.2) we have

$$a \log 2 + b \log 3 \approx \frac{\log 3}{q}(aq + bp).$$

Using the asymptotic calculus (and the hypotheses that D, j are of polylogarithmic size, and the $c_{a,b}$ are of quasipolynomial size) we conclude that the left-hand side of (7.4) is

$$\approx (\frac{\log 3}{q})^j \sum_{0 \leq a,b \leq D} c_{a,b} (ap + bq)^j 2^{an} 3^{bn}. \tag{7.5}$$

The quantity $(\frac{\log 3}{q})^j$ (and its reciprocal) is of quasipolynomial size. Thus, the condition (7.4) is equivalent to the assertion that

$$\sum_{0 \leq a,b \leq D} c_{a,b}(ap+bq)^j 2^{an} 3^{bn} \approx 0$$

for all $0 \leq j \leq J$ and $1 \leq n \leq N$; as the left-hand side is a nonstandard integer, we see from the integrality gap that the condition is in fact equivalent to the *exact* constraint

$$\sum_{0 \leq a,b \leq D} c_{a,b}(ap+bq)^j 2^{an} 3^{bn} = 0 \tag{7.6}$$

for all $0 \leq j \leq J$ and $1 \leq n \leq N$.

Using this reformulation of (7.4), we can now give some upper and lower bounds on the complexity of good polynomials that nearly vanish to a high order on a set Γ_N. We first give an lower bound, that prevents the degree D from being smaller than $N^{1/2}$:

Proposition 7.1.10 (Lower bound)**.** *Let P be a non-trivial good polynomial of degree D that nearly vanishes to order at least 0 on Γ_N. Then $(D+1)^2 > N$.*

Proof. Suppose for contradiction that $(D+1)^2 \leq N$. Then from (7.6) we have

$$\sum_{0 \leq a,b \leq D} c_{a,b}(2^a 3^b)^n = 0$$

for $1 \leq n \leq (D+1)^2$; thus there is a non-trivial linear dependence between the $(D+1)^2$ (nonstandard) vectors $((2^a 3^b)^n)_{1 \leq n \leq (D+1)^2} \in {}^*\mathbf{R}^{(D+1)^2}$ for $0 \leq a,b \leq D$. But, from the formula for the *Vandermonde determinant*, this would imply that two of the $2^a 3^b$ are equal, which is absurd. □

In the converse direction, we can obtain polynomials that vanish to a high order J on Γ_N, but with degree D larger than $N^{1/2}J^{1/2}$:

Proposition 7.1.11 (Upper bound)**.** *Let D, J, N be positive quantities of polylogarithmic size such that*

$$D^2 \ggg NJ.$$

Then there exists a non-trivial good polynomial P of degree at most D that vanishes to order J on Γ_N. Furthermore, P has height at most

$$\exp(O(\frac{NJ^2 \log H}{D^2} + \frac{N^2 J}{D})).$$

Proof. We use the pigeonholing argument of Thue and Siegel. Let M be an positive quantity of quasipolynomial size to be chosen later, and choose coefficients $c_{a,b}$ for $0 \leq a, b \leq D$ that are nonstandard natural numbers between 1 and M. There are $M^{(D+1)^2} \geq \exp(D^2 \log M)$ possible ways to make such a selection. For each such selection, we consider the $N(J+1)$ expressions arising as the left-hand side of (7.6) with $0 \leq j \leq J$ and $1 \leq n \leq N$. These expressions are nonstandard integers whose magnitude is bounded by

$$O((D+1)^2 M O(DH)^J \exp(O(ND)))$$

which by asymptotic calculus simplifies to be bounded by

$$\exp(\log M + O(J \log H) + O(ND)).$$

The number of possible values of these $N(J+1)$ expressions is thus

$$\exp(N(J+1) \log M + O(NJ^2 \log H) + O(N^2 JD)).$$

By the hypothesis $D^2 \gg NJ$ and asymptotic calculus, we can make this quantity less than $\exp(D^2 \log M)$ for some M of size

$$M \ll \exp(O(\frac{NJ^2 \log H}{D^2} + \frac{N^2 J}{D})).$$

In particular, M can be taken to be of polylogarithmic size. Thus, by the pigeonhole principle, one can find two choices for the coefficients $c_{a,b}$ which give equal values for the expressions in the left-hand side of (7.6). Subtracting those two choices we obtain the result. □

7.1.3. The bootstrap. At present, there is no contradiction between the lower bound in Proposition 7.1.10 and the upper bound in Proposition 7.1.11, because there is plenty of room between the two bounds. To bridge the gap between the bounds, we need a bootstrap argument that uses vanishing on one Γ_N to imply vanishing (to slightly lower order) on a larger $\Gamma_{N'}$. The key bootstrap in this regard is:

Proposition 7.1.12 (Bootstrap)**.** *Let D, J, N be unbounded polylogarithmic quantities, such that*

$$N \ggg \log H.$$

Let P be a good polynomial of degree at most D and height $\exp(O(NJ))$, that nearly vanishes to order $2J$ on Γ_N. Then P also vanishes to order J on $\Gamma_{N'}$ for any $N' = o(\frac{J}{D}N)$.

Proof. It is convenient to use complex analysis methods. We consider the entire function

$$f(z) := P(2^z, 3^z),$$

thus by (7.3)

$$f(z) = \sum_{0 \leq a,b \leq D} c_{a,b} 2^{az} 3^{bz}.$$

By hypothesis, we have

$$f^{(j)}(n) \approx 0$$

for all $0 \leq j \leq 2J$ and $1 \leq n \leq N$. We wish to show that

$$f^{(j)}(n') \approx 0$$

for $0 \leq j \leq J$ and $1 \leq n' \leq N'$. Clearly we may assume that $N' \geq n' > N$.

Fix $0 \leq j \leq J$ and $1 \leq n' \leq N'$. To estimate $f^{(j)}(n')$, we consider the contour integral

$$\frac{1}{2\pi i} \int_{|z|=R} \frac{f^{(j)}(z)}{\prod_{n=1}^{N} (z-n)^J} \frac{dz}{z-n'} \tag{7.7}$$

(oriented anticlockwise), where $R \geq 2N'$ is to be chosen later, and estimate it in two different ways. Firstly, we have

$$f^{(j)}(z) = \sum_{0 \leq a,b \leq D} c_{a,b} (a \log 2 + b \log 3)^j 2^{az} 3^{bz},$$

so for $|z| = 2N'$, we have the bound

$$|f^{(j)}(z)| \ll D^2 \exp(o(NJ)) O(D)^J \exp(O(DR))$$

when $|z| = 2N'$, which by the hypotheses and asymptotic calculus, simplifies to

$$|f^{(j)}(z)| \ll \exp(O(NJ + DR)).$$

Also, when $|z| = R$ we have

$$|\prod_{n=1}^{N} (z-n)^J| \geq (R/2)^{NJ}.$$

We conclude the upper bound

$$\exp(O(NJ + DR) - NJ \log R)$$

for the magnitude of (7.7). On the other hand, we can evaluate (7.7) using the *residue theorem*. The integrand has poles at $1, \ldots, N$ and at n'. The simple pole at n' has residue

$$\frac{f^{(j)}(n')}{\prod_{n=1}^{N} (n'-n)^J}.$$

Now we consider the poles at $n = 1, \ldots, N$. For each such n, we see that the first J derivatives of $f^{(j)}$ are quasiexponentially small at n. Thus, by Taylor expansion (and asymptotic calculus), one can express $f^{(j)}(z)$ as the sum of a polynomial of degree J with quasiexponentially small coefficients, plus an entire function that vanishes to order J at n. The latter term contributes

nothing to the residue at n, while from the Cauchy integral formula (applied, for instance, to a circle of radius $1/2$ around n) and asymptotic calculus, we see that the former term contributes a residue is quasiexponentially small. In particular, it is less than $\exp(O(NJ) - NJ \log R)$. We conclude that

$$|\frac{f^{(j)}(n')}{\prod_{n=1}^{N}(n'-n)^J}| \ll \exp(O(NJ + DR) - NJ \log R).$$

We have

$$|\prod_{n=1}^{N}(n'-n)^J| \leq (N')^{NJ}$$

and thus

$$|f^{(j)}(n')| \ll \exp(O(NJ + DR) - NJ \log \frac{R}{N'});$$

choosing R to be a large standard multiple of N' and using the hypothesis $N' = o(\frac{J}{D}N)$, we can simplify this to

$$|f^{(j)}(n')| \ll \exp(-NJ).$$

To improve this bound, we use the integrality gap. Recall that from (7.5) that

$$|f^{(j)}(n')| \approx (\frac{\log 3}{q})^j \sum_{0 \leq a,b \leq D} c_{a,b}(ap+bq)^j 2^{an'} 3^{bn'};$$

in particular, $(\frac{q}{\log 3})^j f^{(j)}(n')$ is quasiexponentially close to a (nonstandard) integer. Since

$$(\frac{q}{\log 3})^j = \exp(O(J \log H)),$$

we have

$$|(\frac{q}{\log 3})^j f^{(j)}(n')| \leq \frac{1}{2}$$

(say). Using the integrality gap, we conclude that

$$(\frac{q}{\log 3})^j f^{(j)}(n') \approx 0$$

which implies that f nearly vanishes to order J on $\Gamma_{N'}$, as required. □

Now we can finish the proof of Proposition 7.1.7 (and hence Proposition 7.1.6). We select quantities D, J, N_0 of polylogarithmic size obeying the bounds

$$\log H \lll N_0 \lll D \lll J$$

and

$$N_0 J \lll D^2,$$

with a gap of a positive power of $\log H$ between each such inequality. For instance, one could take

$$N_0 := \log^2 H$$
$$D := \log^4 H$$
$$J := \log^5 H;$$

many other choices are possible (and one can optimise these choices eventually to get a good value of exponent C' in Proposition 7.1.6).

Using Proposition 7.1.11, we can find a good polynomial P which vanishes to order J on Γ_{N_0}, of height $\exp(O(\frac{N_0 J^2 \log H}{D^2} + \frac{N_0^2 J}{D}))$, and hence (by the assumptions on N_0, D, J) of height $\exp(O(N_0 J))$.

Applying Proposition 7.1.12, P nearly vanishes to order $J/2$ on Γ_{N_1} for any $N_1 = o(\frac{J}{D} N_0)$. Iterating this, an easy induction shows that for any standard $k \geq 1$, P nearly vanishes to order $J/2^k$ on Γ_{N_k} for any $N_k = o((\frac{J}{D})^k N_0)$. As J/D was chosen to be larger than a positive power of $\log H$, we conclude that P nearly vanishes to order at least 0 on Γ_N for any N of polylogarithmic size. But for N large enough, this contradicts Proposition 7.1.10.

Remark 7.1.13. The above argument places a lower bound on quantities such as

$$q \log 2 - p \log 3$$

for integer p, q. *Baker's theorem*, in its full generality, gives a lower bound on quantities such as

$$\beta_0 + \beta_1 \log \alpha_1 + \ldots + \beta_n \log \alpha_n$$

for algebraic numbers $\beta_0, \ldots, \beta_n, \alpha_1, \ldots, \alpha_n$, which is polynomial in the height of the quantities involved, assuming of course that $1, \alpha_1, \ldots, \alpha_n$ are multiplicatively independent, and that all quantities are of bounded degree. The proof is more intricate than the one given above, but follows a broadly similar strategy, and the constants are completely effective.

7.2. The Collatz conjecture, Littlewood-Offord theory, and powers of 2 and 3

One of the most notorious problems in elementary mathematics that remains unsolved is the *Collatz conjecture*, concerning the function $f_0 : \mathbf{N} \to \mathbf{N}$ defined by setting $f_0(n) := 3n + 1$ when n is odd, and $f_0(n) := n/2$ when n is even. (Here, $\mathbf{N}$ is understood to be the positive natural numbers $\{1, 2, 3, \ldots\}$.)

Conjecture 7.2.1 (Collatz conjecture)**.** *For any given natural number n, the orbit $n, f_0(n), f_0^2(n), f_0^3(n), \ldots$ passes through 1 (i.e. $f_0^k(n) = 1$ for some k).*

Open questions with this level of notoriety can lead to what Richard Lipton calls[3] "mathematical diseases". Nevertheless, it can still be diverting to spend a day or two each year on these sorts of questions, before returning to other matters; so I recently had a go at the problem. Needless to say, I didn't solve the problem, but I have a better appreciation of why the conjecture is (a) plausible, and (b) unlikely be proven by current technology, and I thought I would share what I had found out here.

Let me begin with some very well known facts. If n is odd, then $f_0(n) = 3n+1$ is even, and so $f_0^2(n) = \frac{3n+1}{2}$. Because of this, one could replace f_0 by the function $f_1 : \mathbf{N} \to \mathbf{N}$, defined by $f_1(n) = \frac{3n+1}{2}$ when n is odd, and $f_1(n) = n/2$ when n is even, and obtain an equivalent conjecture. Now we see that if one chooses n "at random", in the sense that it is odd with probability $1/2$ and even with probability $1/2$, then f_1 increases n by a factor of roughly $3/2$ half the time, and decreases it by a factor of $1/2$ half the time. Furthermore, if n is uniformly distributed modulo 4, one easily verifies that $f_1(n)$ is uniformly distributed modulo 2, and so $f_1^2(n)$ should be roughly $3/2$ times as large as $f_1(n)$ half the time, and roughly $1/2$ times as large as $f_1(n)$ the other half of the time. Continuing this at a heuristic level, we expect generically that $f_1^{k+1}(n) \approx \frac{3}{2} f_1^k(n)$ half the time, and $f_1^{k+1}(n) \approx \frac{1}{2} f_1^k(n)$ the other half of the time. The logarithm $\log f_1^k(n)$ of this orbit can then be modeled heuristically by a random walk with steps $\log \frac{3}{2}$ and $\log \frac{1}{2}$ occuring with equal probability. The expectation

$$\frac{1}{2} \log \frac{3}{2} + \frac{1}{2} \log \frac{1}{2} = \frac{1}{2} \log \frac{3}{4}$$

is negative, and so (by the classic *gambler's ruin*) we expect the orbit to decrease over the long term. This can be viewed as heuristic justification of the Collatz conjecture, at least in the "average case" scenario in which n is chosen uniform at random (e.g. in some large interval $\{1, \ldots, N\}$). (It also suggests that if one modifies the problem, e.g. by replacing $3n+1$ to $5n+1$, then one can obtain orbits that tend to increase over time, and indeed numerically for this variant one sees orbits that appear to escape to infinity.) Unfortunately, one can only rigorously keep the orbit uniformly distributed modulo 2 for time about $O(\log N)$ or so; after that, the system is too complicated for naive methods to control at anything other than a heuristic level.

Remark 7.2.2. One can obtain a rigorous analogue of the above arguments by extending f_1 from the integers $\mathbf{Z}$ to the 2-*adics* $\mathbf{Z}_2$ (the inverse limit of the cyclic groups $\mathbf{Z}/2^n\mathbf{Z}$). This compact abelian group comes with a *Haar probability measure*, and one can verify that this measure is invariant with respect to f_1; with a bit more effort one can verify that it is ergodic. This

[3] See `rjlipton.wordpress.com/2009/11/04/on-mathematical-diseases`.

suggests the introduction of ergodic theory methods. For instance, using the *pointwise ergodic theorem*, we see that if n is a random 2-*adic* integer, then almost surely the orbit $n, f_1(n), f_1^2(n), \ldots$ will be even half the time and odd half the time asymptotically, thus supporting the above heuristics. Unfortunately, this does not directly tell us much about the dynamics on $\mathbf{Z}$, as this is a measure zero subset of $\mathbf{Z}_2$. More generally, unless a dynamical system is somehow "polynomial", "nilpotent", or "unipotent" in nature, the current state of ergodic theory is usually only able to say something meaningful about *generic* orbits, but not about *all* orbits. For instance, the very simple system $x \to 10x$ on the unit circle $\mathbf{R}/\mathbf{Z}$ is well understood from ergodic theory (in particular, almost all orbits will be uniformly distributed), but the orbit of a specific point, e.g. $\pi \bmod 1$, is still nearly impossible to understand (this particular problem being equivalent to the notorious unsolved question of whether the digits of π are uniformly distributed).

The above heuristic argument only suggests decreasing orbits for *almost all* n (though even this remains unproven, the state of the art is that the number of n in $\{1, \ldots, N\}$ that eventually go to 1 is $\gg N^{0.84}$, see [**KrLa2003**]). It leaves open the possibility of some very rare exceptional n for which the orbit goes to infinity, or gets trapped in a periodic loop. Since the only loop that 1 lies in is $1, 4, 2$ (for f_0) or $1, 2$ (for f_1), we thus may isolate a weaker consequence of the Collatz conjecture:

Conjecture 7.2.3 (Weak Collatz conjecture)**.** *Suppose that n is a natural number such that $f_0^k(n) = n$ for some $k \geq 1$. Then n is equal to 1, 2, or 4.*

Of course, we may replace f_0 with f_1 (and delete "4") and obtain an equivalent conjecture.

This weaker version of the Collatz conjecture is also unproven. However, it was observed in [**BoSo1978**]by Bohm and Sontacchi that this weak conjecture is equivalent to a divisibility problem involving powers of 2 and 3:

Conjecture 7.2.4 (Reformulated weak Collatz conjecture)**.** *There does not exist $k \geq 1$ and integers*

$$0 = a_1 < a_2 < \ldots < a_{k+1}$$

such that $2^{a_{k+1}} - 3^k$ is a positive integer that is a proper divisor of

$$3^{k-1}2^{a_1} + 3^{k-2}2^{a_2} + \ldots + 2^{a_k},$$

i.e.

$$(2^{a_{k+1}} - 3^k)n = 3^{k-1}2^{a_1} + 3^{k-2}2^{a_2} + \ldots + 2^{a_k} \tag{7.8}$$

for some natural number $n > 1$.

Proposition 7.2.5. *Conjecture 7.2.3 and Conjecture 7.2.4 are equivalent.*

Proof. To see this, it is convenient to reformulate Conjecture 7.2.3 slightly. Define an equivalence relation $\sim$ on $\mathbf{N}$ by declaring $a \sim b$ if $a/b = 2^m$ for some integer m, thus giving rise to the quotient space $\mathbf{N}/\sim$ of equivalence classes $[n]$ (which can be placed, if one wishes, in one-to-one correspondence with the odd natural numbers). We can then define a function $f_2 : \mathbf{N}/\sim \to \mathbf{N}/\sim$ by declaring

$$f_2([n]) := [3n + 2^a] \tag{7.9}$$

for any $n \in \mathbf{N}$, where 2^a is the largest power of 2 that divides n. It is easy to see that f_2 is well-defined (it is essentially the *Syracuse function*, after identifying $\mathbf{N}/\sim$ with the odd natural numbers), and that periodic orbits of f_2 correspond to periodic orbits of f_1 or f_0. Thus, Conjecture 7.2.3 is equivalent to the conjecture that [1] is the only periodic orbit of f_2.

Now suppose that Conjecture 7.2.3 failed, thus there exists $[n] \neq [1]$ such that $f_2^k([n]) = [n]$ for some $k \geq 1$. Without loss of generality we may take n to be odd, then $n > 1$. It is easy to see that [1] is the only fixed point of f_2, and so $k > 1$. An easy induction using (7.9) shows that

$$f_2^k([n]) = [3^k n + 3^{k-1}2^{a_1} + 3^{k-2}2^{a_2} + \ldots + 2^{a_k}]$$

where, for each $1 \leq i \leq k$, 2^{a_i} is the largest power of 2 that divides

$$n_i := 3^{i-1}n + 3^{i-2}2^{a_1} + \ldots + 2^{a_{i-1}}. \tag{7.10}$$

In particular, as $n_1 = n$ is odd, $a_1 = 0$. Using the recursion

$$n_{i+1} = 3n_i + 2^{a_i}, \tag{7.11}$$

we see from induction that 2^{a_i+1} divides n_{i+1}, and thus $a_{i+1} > a_i$:

$$0 = a_1 < a_2 < \ldots < a_k.$$

Since $f_2^k([n]) = [n]$, we have

$$2^{a_{k+1}}n = 3^k n + 3^{k-1}2^{a_1} + 3^{k-2}2^{a_2} + \ldots + 2^{a_k} = 3n_k + 2^{a_k}$$

for some integer a_{k+1}. Since $3n_k + 2^{a_k}$ is divisible by 2^{a_k+1}, and n is odd, we conclude $a_{k+1} > a_k$; if we rearrange the above equation as (7.8), then we obtain a counterexample to Conjecture 7.2.4.

Conversely, suppose that Conjecture 7.2.4 failed. Then we have $k \geq 1$, integers

$$0 = a_1 < a_2 < \ldots < a_{k+1}$$

and a natural number $n > 1$ such that (7.8) holds. As $a_1 = 0$, we see that the right-hand side of (7.8) is odd, so n is odd also. If we then introduce

the natural numbers n_i by the formula (7.10), then an easy induction using (7.11) shows that

$$(2^{a_{k+1}} - 3^k)n_i = 3^{k-1}2^{a_i} + 3^{k-2}2^{a_{i+1}} + \ldots + 2^{a_{i+k-1}} \tag{7.12}$$

with the periodic convention $a_{k+j} := a_j + a_{k+1}$ for $j > 1$. As the a_i are increasing in i (even for $i \geq k+1$), we see that 2^{a_i} is the largest power of 2 that divides the right-hand side of (7.12); as $2^{a_{k+1}} - 3^k$ is odd, we conclude that 2^{a_i} is also the largest power of 2 that divides n_i. We conclude that

$$f_2([n_i]) = [3n_i + 2^{a_i}] = [n_{i+1}]$$

and thus $[n]$ is a periodic orbit of f_2. Since n is an odd number larger than 1, this contradicts Conjecture 7.2.4. □

Call a *counterexample* a tuple $(k, a_1, \ldots, a_{k+1})$ that contradicts Conjecture 7.2.4, i.e. an integer $k \geq 1$ and an increasing set of integers

$$0 = a_1 < a_2 < \ldots < a_{k+1}$$

such that (7.8) holds for some $n \geq 1$. We record a simple bound on such counterexamples, due to Terras [**Te1976**] and Garner [**Ga1981**]:

Lemma 7.2.6 (Exponent bounds)**.** *Let $N \geq 1$, and suppose that the Collatz conjecture is true for all $n < N$. Let $(k, a_1, \ldots, a_{k+1})$ be a counterexample. Then*

$$\frac{\log 3}{\log 2}k < a_{k+1} < \frac{\log(3 + \frac{1}{N})}{\log 2}k.$$

Proof. The first bound is immediate from the positivity of $2^{a_{k+1}} - 3^k$. To prove the second bound, observe from the proof of Proposition 7.2.5 that the counterexample $(k, a_1, \ldots, a_{k+1})$ will generate a counterexample to Conjecture 7.2.3, i.e. a non-trivial periodic orbit $n, f(n), \ldots, f^K(n) = n$. As the conjecture is true for all $n < N$, all terms in this orbit must be at least N. An inspection of the proof of Proposition 7.2.5 reveals that this orbit consists of k steps of the form $x \mapsto 3x+1$, and a_{k+1} steps of the form $x \mapsto x/2$. As all terms are at least n, the former steps can increase magnitude by a multiplicative factor of at most $3 + \frac{1}{N}$. As the orbit returns to where it started, we conclude that

$$1 \leq (3 + \frac{1}{N})^k (\frac{1}{2})^{a_{k+1}}$$

whence the claim. □

The Collatz conjecture has already been verified for many values[4] of n. Inserting this into the above lemma, one can get lower bounds on k. For

[4] According to `http://www.ieeta.pt/ tos/3x+1.html`, the conjecture has been verified up to at least $N = 5 \times 10^{18}$.

instance, by methods such as this, it is known that any non-trivial periodic orbit has length at least $105,000$, as shown in [**Ga1981**] (and this bound, which uses the much smaller value $N = 2 \times 10^9$ that was available in 1981, can surely be improved using the most recent computational bounds).

Now we can perform a heuristic count on the number of counterexamples. If we fix k and $a := a_{k+1}$, then $2^a > 3^k$, and from basic combinatorics we see that there are $\binom{a-1}{k-1}$ different ways to choose the remaining integers

$$0 = a_1 < a_2 < \ldots < a_{k+1}$$

to form a potential counterexample $(k, a_1, \ldots, a_{k+1})$. As a crude heuristic, one expects that for a "random" such choice of integers, the expression (7.8) has a probability $1/q$ of holding for some integer n. (Note that q is not divisible by 2 or 3, and so one does not expect the special structure of the right-hand side of (7.8) with respect to those moduli to be relevant. There will be some choices of $a_1, \ldots, a_k$ where the right-hand side in (7.8) is too small to be divisible by q, but using the estimates in Lemma 7.2.6, one expects this to occur very infrequently.) Thus, the total expected number of solutions for this choice of a, k is

$$\frac{1}{q}\binom{a-1}{k-1}.$$

The heuristic number of solutions overall is then expected to be

$$\sum_{a,k} \frac{1}{q}\binom{a-1}{k-1}, \tag{7.13}$$

where, in view of Lemma 7.2.6, one should restrict the double summation to the heuristic regime $a \approx \frac{\log 3}{\log 2} k$, with the approximation here accurate to many decimal places.

We need a lower bound on q. Here, we will use *Baker's theorem* (as discussed in Section 7.1), which among other things gives the lower bound

$$q = 2^a - 3^k \gg 2^a / a^C \tag{7.14}$$

for some absolute constant C. Meanwhile, *Stirling's formula* (as discussed for instance in [**Ta2011c**, §1.2]) combined with the approximation $k \approx \frac{\log 2}{\log 3} a$ gives

$$\binom{a-1}{k-1} \approx \exp(h(\frac{\log 2}{\log 3}))^a$$

where h is the *entropy function*

$$h(x) := -x \log x - (1-x) \log(1-x).$$

A brief computation shows that

$$\exp(h(\frac{\log 2}{\log 3})) \approx 1.9318\ldots$$

and so (ignoring all subexponential terms)

$$\frac{1}{q}\binom{a-1}{k-1} \approx (0.9659\ldots)^a$$

which makes the series (7.13) convergent. (Actually, one does not need the full strength of Lemma 7.2.6 here; anything that kept k well away from $a/2$ would suffice. In particular, one does not need an enormous value of N; even $N = 5$ (say) would be more than sufficient to obtain the heuristic that there are finitely many counterexamples.) Heuristically applying the *Borel-Cantelli lemma*, we thus expect that there are only a finite number of counterexamples to the weak Collatz conjecture (and inserting a bound such as $k \geq 105{,}000$, one in fact expects it to be extremely likely that there are no counterexamples at all).

This, of course, is far short of any rigorous proof of Conjecture 7.2.3. In order to make rigorous progress on this conjecture, it seems that one would need to somehow exploit the structural properties of numbers of the form

$$3^{k-1}2^{a_1} + 3^{k-2}2^{a_2} + \ldots + 2^{a_k}. \tag{7.15}$$

In some very special cases, this can be done. For instance, suppose that one had $a_{i+1} = a_i + 1$ with at most one exception (this is essentially what is called a 1-*cycle* in [**St1978**]). Then (7.15) simplifies via the geometric series formula to a combination of just a bounded number of powers of 2 and 3, rather than an unbounded number. In that case, one can start using tools from transcendence theory such as Baker's theorem to obtain good results; for instance, in [**St1978**], it was shown that 1-cycles cannot actually occur, and similar methods have been used to show that m-cycles (in which there are at most m exceptions to $a_{i+1} = a_i + 1$) do not occur for any $m \leq 63$, as was shown in [**Side2005**]. However, for general increasing tuples of integers $a_1, \ldots, a_k$, there is no such representation by bounded numbers of powers, and it does not seem that methods from transcendence theory will be sufficient to control the expressions (7.15) to the extent that one can understand their divisibility properties by quantities such as $2^a - 3^k$.

Amusingly, there is a slight connection to *Littlewood-Offord theory* in additive combinatorics - the study of the 2^n random sums

$$\pm v_1 \pm v_2 \pm \ldots \pm v_n$$

generated by some elements $v_1, \ldots, v_n$ of an additive group G, or equivalently, the vertices of an n-dimensional parallelepiped inside G. Here, the relevant group is $\mathbf{Z}/q\mathbf{Z}$. The point is that if one fixes k and a_{k+1} (and hence q), and lets $a_1, \ldots, a_k$ vary inside the simplex

$$\Delta := \{(a_1, \ldots, a_k) \in \mathbf{N}^k : 0 = a_1 < \ldots < a_k < a_{k+1}\}$$

then the set S of all sums[5] of the form (7.15) (viewed as an element of $\mathbf{Z}/q\mathbf{Z}$) contains many large parallelepipeds. This is because the simplex Δ contains many large cubes. Indeed, if one picks a typical element $(a_1, \ldots, a_k)$ of Δ, then one expects (thanks to Lemma 7.2.6) that there there will be $\gg k$ indices $1 \leq i_1 < \ldots < i_m \leq k$ such that $a_{i_j+1} > a_{i_j} + 1$ for $j = 1, \ldots, m$, which allows one to adjust each of the a_{i_j} independently by 1 if desired and still remain inside Δ. This gives a cube in Δ of dimension $\gg k$, which then induces a parallelepiped of the same dimension in S. A short computation shows that the generators of this parallelepiped consist of products of a power of 2 and a power of 3, and in particular will be coprime to q.

If the weak Collatz conjecture is true, then the set S must avoid the residue class 0 in $\mathbf{Z}/q\mathbf{Z}$. Let us suppose temporarily that we did not know about Baker's theorem (and the associated bound (7.14)), so that q could potentially be quite small. Then we would have a large parallelepiped inside a small cyclic group $\mathbf{Z}/q\mathbf{Z}$ that did not cover all of $\mathbf{Z}/q\mathbf{Z}$, which would not be possible for q small enough. Indeed, an easy induction shows that a d-dimensional parallelepiped in $\mathbf{Z}/q\mathbf{Z}$, with all generators coprime to q, has cardinality at least $\min(q, d+1)$. This argument already shows the lower bound $q \gg k$. In other words, we have

Proposition 7.2.7. *Suppose the weak Collatz conjecture is true. Then for any natural numbers a, k with $2^a > 3^k$, one has $2^a - 3^k \gg k$.*

This bound is very weak when compared against the unconditional bound (7.14). However, I know of no way to get a nontrivial separation property between powers of 2 and powers of 3 other than via transcendence theory methods. Thus, this result strongly suggests that any proof of the Collatz conjecture must either use existing results in transcendence theory, or else must contribute a new method to give non-trivial results in transcendence theory. (This already rules out a lot of possible approaches to solve the Collatz conjecture.)

By using more sophisticated tools in additive combinatorics, one can improve the above proposition (though it is still well short of the transcendence theory bound (7.14)):

Proposition 7.2.8. *Suppose the weak Collatz conjecture is true. Then for any natural numbers a, k with $2^a > 3^k$, one has $2^a - 3^k \gg (1+\varepsilon)^k$ for some absolute constant $\varepsilon > 0$.*

Proof. (Informal sketch only) Suppose not, then we can find a, k with $q := 2^a - 3^k$ of size $(1 + o(1))^k = \exp(o(k))$. We form the set S as before, which

[5]Note, incidentally, that once one fixes k, all the sums of the form (7.15) are distinct; because given (7.15) and k, one can read off 2^{a_1} as the largest power of 2 that divides (7.15), and then subtracting off $3^{k-1}2^{a_1}$ one can then read off 2^{a_2}, and so forth.

contains parallelepipeds in $\mathbf{Z}/q\mathbf{Z}$ of large dimension $d \gg k$ that avoid 0. We can count the number of times 0 occurs in one of these parallelepipeds by a standard Fourier-analytic computation involving Riesz products (see [**TaVu2006**, Chapter 7] or [**Ma2010**]). Using this Fourier representation, the fact that this parallelepiped avoids 0 (and the fact that $q = \exp(o(k)) = \exp(o(d))$) forces the generators $v_1, \ldots, v_d$ to be concentrated in a *Bohr set*, in that one can find a non-zero frequency $\xi \in \mathbf{Z}/q\mathbf{Z}$ such that $(1 - o(1))d$ of the d generators lie in the set $\{v : \xi v = o(q) \bmod q\}$. However, one can choose the generators to essentially have the structure of a (generalised) geometric progression (up to scaling, it resembles something like $2^i 3^{\lfloor \alpha i \rfloor}$ for i ranging over a generalised arithmetic progression, and α a fixed irrational), and one can show that such progressions cannot be concentrated in Bohr sets (this is similar in spirit to the exponential sum estimates of Bourgain [**Bo2005**] on approximate multiplicative subgroups of $\mathbf{Z}/q\mathbf{Z}$, though one can use more elementary methods here due to the very strong nature of the Bohr set concentration (being of the "99% concentration" variety rather than the "1% concentration").). This furnishes the required contradiction. □

Thus we see that any proposed proof of the Collatz conjecture must either use transcendence theory, or introduce new techniques that are powerful enough to create exponential separation between powers of 2 and powers of 3.

Unfortunately, once one uses the transcendence theory bound (7.14), the size q of the cyclic group $\mathbf{Z}/q\mathbf{Z}$ becomes larger than the volume of any cube in S, and Littlewood-Offord techniques are no longer of much use (they can be used to show that S is highly equidistributed in $\mathbf{Z}/q\mathbf{Z}$, but this does not directly give any way to prevent S from containing 0).

One possible toy model problem for the (weak) Collatz conjecture is a conjecture of Erdos [**Er1979**] asserting that for $n > 8$, the base 3 representation of 2^n contains at least one 2. (See [**La2009**] for some work on this conjecture and on related problems.) To put it another way, the conjecture asserts that there are no integer solutions to

$$2^n = 3^{a_1} + 3^{a_2} + \ldots + 3^{a_k}$$

with $n > 8$ and $0 \leq a_1 < \ldots < a_k$. (When $n = 8$, of course, one has $2^8 = 3^0 + 3^1 + 3^2 + 3^5$.) In this form we see a resemblance to Conjecture 7.2.4, but it looks like a simpler problem to attack (though one which is still a fair distance beyond what one can do with current technology). Note that one has a similar heuristic support for this conjecture as one does for Proposition 7.2.4; a number of magnitude 2^n has about $n\frac{\log 2}{\log 3}$ base 3 digits, so the heuristic probability that none of these digits are equal to 2 is $3^{-n\frac{\log 2}{\log 3}} = 2^{-n}$, which is absolutely summable.

7.3. Erdos's divisor bound

One of the basic problems in *analytic number theory* is to obtain bounds and asymptotics for sums of the form[6]

$$\sum_{n \leq x} f(n)$$

in the limit $x \to \infty$, where n ranges over natural numbers less than x, and $f : \mathbf{N} \to \mathbf{C}$ is some *arithmetic function* of number-theoretic interest. For instance, the celebrated *prime number theorem* is equivalent to the assertion

$$\sum_{n \leq x} \Lambda(n) = x + o(x)$$

where $\Lambda(n)$ is the *von Mangoldt function* (equal to $\log p$ when n is a power of a prime p, and zero otherwise), while the infamous *Riemann hypothesis* is equivalent to the stronger assertion

$$\sum_{n \leq x} \Lambda(n) = x + O(x^{1/2+o(1)}).$$

It is thus of interest to develop techniques to estimate such sums $\sum_{n \leq x} f(n)$. Of course, the difficulty of this task depends on how "nice" the function f is. The functions f that come up in number theory lie on a broad spectrum of "niceness", with some particularly nice functions being quite easy to sum, and some being insanely difficult.

At the easiest end of the spectrum are those functions f that exhibit some sort of regularity or "smoothness". Examples of smoothness include "Archimedean" smoothness, in which $f(n)$ is the restriction of some smooth function $f : \mathbf{R} \to \mathbf{C}$ from the reals to the natural numbers, and the derivatives of f are well controlled. A typical example is

$$\sum_{n \leq x} \log n.$$

One can already get quite good bounds on this quantity by comparison with the integral $\int_1^x \log t \; dt$, namely

$$\sum_{n \leq x} \log n = x \log x - x + O(\log x),$$

with sharper bounds available by using tools such as the *Euler-Maclaurin formula* (see [**Ta2011d**, §3.7]). Exponentiating such asymptotics, incidentally, leads to one of the standard proofs of *Stirling's formula* (as discussed in [**Ta2011c**, §1.2]).

[6] It is also often convenient to replace this sharply truncated sum with a smoother sum such as $\sum_n f(n)\psi(n/x)$ for some smooth cutoff ψ, but we will not discuss this technicality here.

One can also consider "non-Archimedean" notions of smoothness, such as periodicity relative to a small period q. Indeed, if f is periodic with period q (and is thus essentially a function on the cyclic group $\mathbf{Z}/q\mathbf{Z}$), then one has the easy bound

$$\sum_{n\leq x} f(n) = \frac{x}{q} \sum_{n\in\mathbf{Z}/q\mathbf{Z}} f(n) + O(\sum_{n\in\mathbf{Z}/q\mathbf{Z}} |f(n)|).$$

In particular, we have the fundamental estimate

$$\sum_{n\leq x: q|n} 1 = \frac{x}{q} + O(1). \tag{7.16}$$

This is a good estimate when q is much smaller than x, but as q approaches x in magnitude, the error term $O(1)$ begins to overwhelm the main term $\frac{n}{q}$, and one needs much more delicate information on the fractional part of $\frac{n}{q}$ in order to obtain good estimates at this point.

One can also consider functions f which combine "Archimedean" and "non-Archimedean" smoothness into an "adelic" smoothness. We will not define this term precisely here (though the concept of a *Schwartz-Bruhat function* is one way to capture this sort of concept), but a typical example might be

$$\sum_{n\leq x} \chi(n) \log n$$

where χ is periodic with some small period q. By using techniques such as *summation by parts*, one can estimate such sums using the techniques used to estimate sums of periodic functions or functions with (Archimedean) smoothness.

Another class of functions that is reasonably well controlled are the *multiplicative functions*, in which $f(nm) = f(n)f(m)$ whenever n, m are coprime. Here, one can use the powerful techniques of *multiplicative number theory*, for instance by working with the *Dirichlet series*

$$\sum_{n=1}^{\infty} \frac{f(n)}{n^s}$$

which are clearly related to the partial sums $\sum_{n\leq x} f(n)$ (essentially via the *Mellin transform*, a cousin of the Fourier and Laplace transforms); for this section we ignore the (important) issue of how to make sense of this series when it is not absolutely convergent (but see [**Ta2011d**, §3.7] for more discussion). A primary reason that this technique is effective is that the Dirichlet series of a multiplicative function factorises as an *Euler product*

$$\sum_{n=1}^{\infty} \frac{f(n)}{n^s} = \prod_p (\sum_{j=0}^{\infty} \frac{f(p^j)}{p^{js}}).$$

One also obtains similar types of representations for functions that are not quite multiplicative, but are closely related to multiplicative functions, such as the von Mangoldt function Λ (whose Dirichlet series $\sum_{n=1}^\infty \frac{\Lambda(n)}{n^s} = -\frac{\zeta'(s)}{\zeta(s)}$ is not given by an Euler product, but instead by the *logarithmic derivative* of an Euler product).

Moving another notch along the spectrum between well-controlled and ill-controlled functions, one can consider functions f that are *divisor sums* such as

$$f(n) = \sum_{d \leq R; d|n} g(d) = \sum_{d \leq R} 1_{d|n} g(d)$$

for some other arithmetic function g, and some *level* R. This is a linear combination of periodic functions $1_{d|n} g(d)$ and is thus *technically* periodic in n (with period equal to the least common multiple of all the numbers from 1 to R), but in practice this periodic is far too large to be useful (except for extremely small levels R, e.g. $R = O(\log x)$). Nevertheless, we can still control the sum $\sum_{n \leq x} f(n)$ simply by rearranging the summation:

$$\sum_{n \leq x} f(n) = \sum_{d \leq R} g(d) \sum_{n \leq x: d|n} 1$$

and thus by (7.16) one can bound this by the sum of a main term $x \sum_{d \leq R} \frac{g(d)}{d}$ and an error term $O(\sum_{d \leq R} |g(d)|)$. As long as the level R is significantly less than x, one may expect the main term to dominate, and one can often estimate this term by a variety of techniques (for instance, if g is multiplicative, then multiplicative number theory techniques are quite effective, as mentioned previously). Similarly for other slight variants of divisor sums, such as expressions of the form

$$\sum_{d \leq R; d|n} g(d) \log \frac{n}{d}$$

or expressions of the form

$$\sum_{d \leq R} F_d(n)$$

where each F_d is periodic with period d.

One of the simplest examples of this comes when estimating the *divisor function*

$$\tau(n) := \sum_{d|n} 1,$$

which counts the number of divisors up to n. This is a multiplicative function, and is therefore most efficiently estimated using the techniques of multiplicative number theory; but for reasons that will become clearer later, let

us "forget" the multiplicative structure and estimate the above sum by more elementary methods. By applying the preceding method, we see that

$$\sum_{n\leq x}\tau(n) = \sum_{d\leq x}\sum_{n\leq x:d|n} 1$$
$$= \sum_{d\leq x}(\frac{x}{d}+O(1))$$
$$= x\log x + O(x). \tag{7.17}$$

Here, we are (barely) able to keep the error term smaller than the main term; this is right at the edge of the divisor sum method, because the level R in this case is equal to x. Unfortunately, at this high choice of level, it is not always possible to always keep the error term under control like this. For instance, if one wishes to use the standard divisor sum representation

$$\Lambda(n) = \sum_{d|n}\mu(d)\log\frac{n}{d},$$

where $\mu(n)$ is the *Möbius function* (defined to equal $(-1)^k$ when n is the product of k distinct primes, and zero otherwise), to compute $\sum_{n\leq x}\Lambda(n)$, then one ends up looking at

$$\sum_{n\leq x}\Lambda(n) = \sum_{d\leq x}\mu(d)\sum_{n\leq x:d|n}\log\frac{n}{d}$$
$$= \sum_{d\leq x}\mu(d)(\frac{n}{d}\log\frac{n}{d} - \frac{n}{d} + O(\log\frac{n}{d}))$$

From Dirichlet series methods, it is not difficult to establish the identities

$$\lim_{s\to 1^+}\sum_{n=1}^{\infty}\frac{\mu(n)}{n^s} = 0$$

and

$$\lim_{s\to 1^+}\sum_{n=1}^{\infty}\frac{\mu(n)\log n}{n^s} = -1.$$

This suggests (but does not quite prove) that one has

$$\sum_{n=1}^{\infty}\frac{\mu(n)}{n} = 0 \tag{7.18}$$

and

$$\sum_{n=1}^{\infty}\frac{\mu(n)\log n}{n} = -1 \tag{7.19}$$

in the sense of conditionally convergent series. Assuming one can justify this (which, ultimately, requires one to exclude zeroes of the Riemann zeta

function on the line $\mathrm{Re}(s) = 1$, as discussed in [**Ta2010b**, §1.12]), one is eventually left with the estimate $x + O(x)$, which is useless as a lower bound (and recovers only the classical Chebyshev estimate $\sum_{n \leq x} \Lambda(n) \ll x$ as the upper bound). The inefficiency here when compared to the situation with the divisor function τ can be attributed to the signed nature of the Möbius function $\mu(n)$, which causes some cancellation in the divisor sum expansion that needs to be compensated for with improved estimates.

However, there are a number of tricks available to reduce the level of divisor sums. The simplest comes from exploiting the change of variables $d \mapsto \frac{n}{d}$, which can in principle reduce the level by a square root. For instance, when computing the divisor function $\tau(n) = \sum_{d|n} 1$, one can observe using this change of variables that every divisor of n above $\sqrt{n}$ is paired with one below $\sqrt{n}$, and so we have

$$\tau(n) = 2 \sum_{d \leq \sqrt{n}: d|n} 1 \tag{7.20}$$

except when n is a perfect square, in which case one must subtract one from the right-hand side. Using this reduced-level divisor sum representation, one can obtain an improvement to (7.17), namely

$$\sum_{n \leq x} \tau(n) = x \log x + (2\gamma - 1)x + O(\sqrt{x}).$$

This type of argument is also known as the *Dirichlet hyperbola method*. A variant of this argument can also deduce the prime number theorem from (7.18), (7.19) (and with some additional effort, one can even drop the use of (7.19)).

Using this square root trick, one can now also control divisor sums such as

$$\sum_{n \leq x} \tau(n^2 + 1).$$

(Note that $\tau(n^2 + 1)$ has no multiplicativity properties in n, and so multiplicative number theory techniques cannot be directly applied here.) The level of the divisor sum here is initially of order x^2, which is too large to be useful; but using the square root trick, we can expand this expression as

$$2 \sum_{n \leq x} \sum_{d \leq n: d|n^2+1} 1$$

which one can rewrite as

$$2 \sum_{d \leq x} \sum_{d \leq n \leq x: n^2+1 = 0 \bmod d} 1.$$

The constraint $n^2 + 1 = 0 \bmod d$ is periodic in n with period d, so we can write this as

$$2 \sum_{d \leq x} (\frac{x}{d} \rho(d) + O(\rho(d)))$$

where $\rho(d)$ is the number of solutions in $\mathbf{Z}/d\mathbf{Z}$ to the equation $n^2 + 1 = 0 \bmod d$, and so

$$\sum_{n \leq x} \tau(n^2+1) = 2x \sum_{d \leq x} \frac{\rho(d)}{d} + O(\sum_{d \leq x} \rho(d)).$$

The function ρ is multiplicative, and can be easily computed at primes p and prime powers p^j using tools such as *quadratic reciprocity* and *Hensel's lemma*. For instance, by *Fermat's two-square theorem*, $\rho(p)$ is equal to 2 for $p = 1 \bmod 4$ and 0 for $p = 3 \bmod 4$. From this and standard multiplicative number theory methods (e.g. by obtaining asymptotics on the Dirichlet series $\sum_d \frac{\rho(d)}{d^s}$), one eventually obtains the asymptotic

$$\sum_{d \leq x} \frac{\rho(d)}{d} = \frac{3}{2\pi} \log x + O(1)$$

and also

$$\sum_{d \leq x} \rho(d) = O(x)$$

and thus

$$\sum_{n \leq x} \tau(n^2+1) = \frac{3}{\pi} x \log x + O(x).$$

Similar arguments give asymptotics for τ on other quadratic polynomials; see for instance [**Ho1963**], [**Mc1995**], [**Mc1997**], [**Mc1999**]. Note that the irreducibility of the polynomial will be important. If one considers instead a sum involving a reducible polynomial, such as $\sum_{n \leq x} \tau(n^2-1)$, then the analogous quantity $\rho(n)$ becomes significantly larger, leading to a larger growth rate (of order $x \log^2 x$ rather than $x \log x$) for the sum.

However, the square root trick is insufficient by itself to deal with higher order sums involving the divisor function, such as

$$\sum_{n \leq x} \tau(n^3+1);$$

the level here is initially of order x^3, and the square root trick only lowers this to about $x^{3/2}$, creating an error term that overwhelms the main term. And indeed, the asymptotic for such this sum has not yet been rigorously established (although if one heuristically drops error terms, one can arrive at a reasonable conjecture for this asymptotic), although some results are known if one averages over additional parameters (see e.g. [**Gr1970**], [**Ma2012**].

Nevertheless, there is an ingenious argument of Erdös [**Er1952**] that allows one to obtain good *upper* and *lower* bounds for these sorts of sums, in particular establishing the asymptotic

$$x \log x \ll \sum_{n \leq x} \tau(P(n)) \ll x \log x \tag{7.21}$$

for any *fixed* irreducible non-constant polynomial P that maps $\mathbf{N}$ to $\mathbf{N}$ (with the implied constants depending of course on the choice of P). There is also the related moment bound

$$\sum_{n \leq x} \tau^m(P(n)) \ll x \log^{O(1)} x \tag{7.22}$$

for any fixed P (not necessarily irreducible) and any fixed $m \geq 1$, due to van der Corput [**va1939**]; this bound is in fact used to dispose of some error terms in the proof of (7.21). These should be compared with what one can obtain from the *divisor bound* $\tau(n) \ll n^{O(1/\log\log n)}$ (see [**Ta2009**, §1.6]) and the trivial bound $\tau(n) \geq 1$, giving the bounds

$$x \ll \sum_{n \leq x} \tau^m(P(n)) \ll x^{1+O(\frac{1}{\log\log x})}$$

for any fixed $m \geq 1$.

The lower bound in (7.21) is easy, since one can simply lower the level in (7.20) to obtain the lower bound

$$\tau(n) \geq \sum_{d \leq n^\theta : d|n} 1$$

for any $\theta > 0$, and the preceding methods then easily allow one to obtain the lower bound by taking θ small enough (more precisely, if P has degree d, one should take θ equal to $1/d$ or less). The upper bounds in (7.21) and (7.22) are more difficult. Ideally, if we could obtain upper bounds of the form

$$\tau(n) \ll \sum_{d \leq n^\theta : d|n} 1 \tag{7.23}$$

for any fixed $\theta > 0$, then the preceding methods would easily establish both results. Unfortunately, this bound can fail, as illustrated by the following example. Suppose that n is the product of k distinct primes $p_1 \dots p_k$, each of which is close to $n^{1/k}$. Then n has 2^k divisors, with $\binom{n}{j}$ of them close to $n^{j/k}$ for each $0 \dots j \leq k$. One can think of (the logarithms of) these divisors as being distributed according to what is essentially a *Bernoulli distribution*, thus a randomly selected divisor of n has magnitude about $n^{j/k}$, where j is a random variable which has the same distribution as the number of heads in k independently tossed fair coins. By the law of large numbers, j should

concentrate near $k/2$ when k is large, which implies that the majority of the divisors of n will be close to $n^{1/2}$. Sending $k \to \infty$, one can show that the bound (7.23) fails whenever $\theta < 1/2$.

This however can be fixed in a number of ways. First of all, even when $\theta < 1/2$, one can show weaker substitutes for (7.23). For instance, for any fixed $\theta > 0$ and $m \geq 1$ one can show a bound of the form

$$\tau(n)^m \ll \sum_{d \leq n^\theta : d|n} \tau(d)^C \tag{7.24}$$

for some C depending only on m, θ. This nice elementary inequality (first observed in [**La1989**]) already gives a quite short proof of van der Corput's bound (7.22).

For Erdős's upper bound (7.21), though, one cannot afford to lose these additional factors of $\tau(d)$, and one must argue more carefully. Here, the key observation is that the counterexample discussed earlier - when the natural number n is the product of a large number of fairly small primes - is quite atypical; most numbers have at least one large prime factor. For instance, the number of natural numbers less than x that contain a prime factor between $x^{1/2}$ and x is equal to

$$\sum_{x^{1/2} \leq p \leq x} (\frac{x}{p} + O(1)),$$

which, thanks to *Mertens' theorem*

$$\sum_{p \leq x} \frac{1}{p} = \log\log x + M + o(1)$$

for some absolute constant M, is comparable to x. In a similar spirit, one can show by similarly elementary means that the number of natural numbers m less than x that are $x^{1/m}$*-smooth*, in the sense that all prime factors are at most $x^{1/m}$, is only about $m^{-cm}x$ or so. Because of this, one can hope that the bound (7.23), while not true in full generality, will still be true for *most* natural numbers n, with some slightly weaker substitute available (such as (7.22)) for the exceptional numbers n. This turns out to be the case by an elementary but careful argument.

The Erdős argument is quite robust; for instance, the more general inequality

$$x \log^{2^m-1} x \ll \sum_{n \leq x} \tau(P(n))^m \ll x \log^{2^m-1} x$$

for fixed irreducible P and $m \geq 1$, which improves van der Corput's inequality (7.23) was shown in [**De1971**] using the same methods. (A slight error in the original paper of Erdős was also corrected in this paper.) In

[**ElTa2011**], we also applied this method to obtain bounds such as

$$\sum_{a\leq A}\sum_{b\leq B}\tau(a^2b+1)\ll AB\log(A+B),$$

which turn out to be enough to obtain the right asymptotics for the number of solutions to the equation $\frac{4}{p}=\frac{1}{x}+\frac{1}{y}+\frac{1}{z}$.

7.3.1. Landreau's argument. We now prove (7.24), and use this to show (7.22).

Suppose first that all prime factors of n have magnitude at most $n^{c/2}$. Then by a greedy algorithm, we can factorise n as the product $n=n_1\dots n_r$ of numbers between $n^{c/2}$ and n^c. In particular, the number r of terms in this factorisation is at most $2/c$. By the trivial inequality $\tau(ab)\leq\tau(a)\tau(b)$ we have

$$\tau(n)\leq\tau(n_1)\dots\tau(n_r)$$

and thus by the pigeonhole principle one has

$$\tau(n)^m\leq\tau(n_j)^{2m/c}$$

for some j. Since n_j is a factor of n that is at most n^c, the claim follows in this case (taking $C:=2m/c$).

Now we consider the general case, in which n may contain prime factors that exceed n^c. There are at most $1/c$ such factors (counting multiplicity). Extracting these factors out first and then running the greedy algorithm again, we may factorise $n=n_1\dots n_rq$ where the n_i are as before, and q is the product of at most $1/c$ primes. In particular, $\tau(q)\leq 2^{1/c}$ and thus

$$\tau(n)\leq 2^{1/c}\tau(n_1)\dots\tau(n_r).$$

One now argues as before (conceding a factor of $2^{1/c}$, which is acceptable) to obtain (7.24) in full generality. (Note that this illustrates a useful principle, which is that large prime factors of n are essentially harmless for the purposes of upper bounding $\tau(n)$.)

Now we prove (7.22). From (7.24) we have

$$\tau(P(n))^m\ll\sum_{d\leq x:d|P(n)}\tau(d)^{O(1)}$$

for any $n\leq x$, and hence we can bound $\sum_{n\leq x}\tau(P(n))^m$ by

$$\ll\sum_{d\leq x}\tau(d)^{O(1)}\sum_{n\leq x:d|n;P(n)=0 \bmod d}1.$$

The inner sum is $\frac{x}{d}\rho(d)+O(\rho(d))=O(\frac{x}{d}\rho(d))$, where $\rho(d)$ is the number of roots of P mod d. Now, for fixed P, it is easy to see that $\rho(p)=O(1)$ for all primes p, and from Hensel's lemma one soon extends this to $\rho(p^j)=$

$O(1)$ for all prime powers p. (This is easy when p does not divide the *discriminant* $\Delta(P)$ of p, as the zeroes of P mod p are then simple. There are only finitely many primes that do divide the discriminant, and they can each be handled separately by Hensel's lemma and an induction on the degree of P.) Meanwhile, from the Chinese remainder theorem, ρ is multiplicative. From this we obtain the crude bound $\rho(d) \ll \tau(d)^{O(1)}$, and so we obtain a bound

$$\sum_{n \leq x} \tau(P(n))^m \ll x \sum_{d \leq x} \frac{\tau(d)^{O(1)}}{d}.$$

This sum can easily be bounded by $x \log^{O(1)} x$ by multiplicative number theory techniques, e.g. by first computing the Dirichlet series

$$\sum_{d=1}^{\infty} \frac{\tau(d)^{O(1)}}{d^{1+\frac{1}{\log x}}}$$

via the Euler product. This proves (7.22).

7.3.2. Erdős' argument. Now we prove (7.21). We focus on the upper bound, as the proof of the lower bound has already been sketched.

We first make a convenient observation: from (7.22) (with $m = 2$) and the Cauchy-Schwarz inequality, we see that we have

$$\sum_{n \in E} \tau(P(n)) \ll x \log x$$

whenever E is a subset of the natural numbers less than x of cardinality $O(x \log^{-C} x)$ for some sufficiently large C. Thus we have the freedom to restrict attention to "generic" n, where by "generic" we mean "lying outside of an exceptional set of cardinality $O(x \log^{-C} x)$ for the C specified above".

Let us now look at the behaviour of $P(n)$ for generic n. We first control the total number of prime factors:

Lemma 7.3.1. *For generic $n \leq x$, $P(n)$ has $O(\log \log x)$ distinct prime factors.*

This result is consistent with the Hardy-Ramanujan and Erdős-Kac theorems [**HaRa1917**], [**Ka1940**], though it does not quite follow from these results (because $P(n)$ lives in quite a sparse set of natural numbers).

Proof. If $P(n)$ has more than $A \log_2 \log x$ prime factors for some A, then $P(n)$ has at least $\log^A x$ divisors, thus $\tau(P(n)) \geq \log^A X$. The claim then follows from (7.22) (with $m = 1$) and Markov's inequality, taking A large enough. □

Next, we try to prevent repeated prime factors:

Lemma 7.3.2. *For generic $n \leq x$, the prime factors of $P(n)$ between $\log^C x$ and $x^{1/2}$ are all distinct.*

Proof. If p is a prime between $\log^C x$ and $x^{1/2}$, then the total number of $n \leq x$ for which p^2 divides $P(n)$ is

$$\rho(p^2)\frac{x}{p^2} + O(\rho(p^2)) = O(\frac{x}{p^2}),$$

so the total number of x that fail the above property is

$$\ll \sum_{\log^C x \leq p \leq x^{1/2}} \frac{x}{p^2} \ll \frac{x}{\log^C x}$$

which is acceptable. □

It is difficult to increase the upper bound here beyond $x^{1/2}$, but fortunately we will not need to go above this bound. The lower bound cannot be significantly reduced; for instance, it is quite likely that $P(n)$ will be divisible by 2^2 for a positive fraction of n. But we have the following substitute:

Lemma 7.3.3. *For generic $n \leq x$, there are no prime powers p^j dividing $P(n)$ with $p < x^{1/(\log\log x)^2}$ and $p^j \geq x^{1/(\log\log x)^2}$.*

Proof. By the preceding lemma, we can restrict attention to primes p with $p < \log^C x$. For each such p, let p^j be the first power of p exceeding $x^{1/(\log\log x)^2}$. Arguing as before, the total number of $n \leq x$ for which p^j divides $P(n)$ is

$$\ll \frac{x}{p^j} \ll \frac{x}{x^{1/(\log\log x)^2}};$$

on the other hand, there are at most $\log^C x$ primes p to consider. The claim then follows from the union bound. □

We now have enough information on the prime factorisation of $P(n)$ to proceed. We arrange the prime factors of $P(n)$ in increasing order (allowing repetitions):

$$P(n) = p_1 \dots p_J.$$

Let $0 \leq j \leq J$ be the largest integer for which $p_1 \dots p_j \leq x$. Suppose first that $J = j + O(1)$, then as in the previous section we would have

$$\tau(P(n)) \ll \tau(p_1 \dots p_j) \leq \sum_{d \leq x : d | P(n)} 1$$

which is an estimate of the form (7.23), and thus presumably advantageous.

Now suppose that J is much larger than j. Since $P(n) = O(x^{O(1)})$, this implies in particular that $p_{j+1} \leq x^{1/2}$ (say), which forces

$$x^{1/2} \leq p_1 \dots p_j \leq x \tag{7.25}$$

and $p_j \leq x^{1/2}$.

For generic n, we have at most $O(\log\log x)$ distinct prime factors, and each such distinct prime less than $x^{1/(\log\log x)^2}$ contributes at most $x^{1/(\log\log x)^2}$ to the product $p_1 \dots p_j$. We conclude that generically, at least one of these primes $p_1, \dots, p_j$ must exceed $x^{1/(\log\log x)^2}$, thus we generically have

$$x^{1/(\log\log x)^2} \leq p_j \leq x^{1/2}.$$

In particular, we have

$$x^{1/(r+1)} \leq p_j \leq x^{1/r}$$

for some $2 \leq r \leq (\log\log x)^2$. This makes the quantity $p_1 \dots p_j$ $x^{1/r}$*-smooth*, i.e. all the prime factors are at most $x^{1/r}$. On the other hand, the remaining prime factors $p_{j+1}, \dots, p_J$ are at least $x^{1/(r+1)}$, and $P(n) = O(x^{O(1)})$, so we have $J = j + O(r)$. Thus we can write $P(n)$ as the product of $p_1 \dots p_j$ and at most $O(r)$ additional primes, which implies that

$$\begin{aligned}\tau(P(n)) &\ll \exp(O(r))\tau(p_1 \dots p_j) \\ &= \exp(O(r)) \sum_{d : d | p_1 \dots p_j} 1.\end{aligned}$$

The exponential factor looks bad, but we can offset it by the $x^{1/r}$-smooth nature of $p_1 \dots p_j$, which is inherited by its factors d. From (7.25), d is at most x; by using the square root trick, we can restrict d to be *at least* the square root of $p_1 \dots p_j$, and thus to be at least $x^{1/4}$. Also, d divides $P(n)$, and as such inherits many of the prime factorisation properties of $P(n)$; in particular, $O(\log\log x)$ distinct prime factors, and d has no prime powers p^j dividing d with $p < x^{1/(\log\log x)^2}$ and $p^j \geq x^{1/(\log\log x)^2}$.

To summarise, we have shown the following variant of (7.23):

Lemma 7.3.4 (Lowering the level)**.** *For generic $n \leq x$, we*

$$\tau(P(n)) \ll \exp(O(r)) \sum_{d \in S_r : d | P(n)} 1$$

for some $1 \leq r \leq (\log\log x)^2$, where S_r is the set of all $x^{1/r}$-smooth numbers d between $x^{1/4}$ and x with $O(\log\log x)$ distinct prime factors, and such that there are no prime powers p^j dividing d with $p < x^{1/(\log\log x)^2}$ and $p^j \geq x^{1/(\log\log x)^2}$.

Applying this lemma (and discarding the non-generic n), we can thus upper bound $\sum_{n \leq x} \tau(P(n))$ (up to acceptable errors) by

$$\ll \sum_{1 \leq r \leq (\log\log x)^2} \exp(O(r)) \sum_{n \leq x} \sum_{d \in S_r : d | P(n)} 1.$$

The level is now less than x and we can use the usual methods to estimate the inner sums:

$$\sum_{n\leq x}\sum_{d\in S_r:d|P(n)} 1 \ll x\sum_{d\in S_r}\frac{\rho(d)}{d}.$$

Thus it suffices to show that

$$\sum_{1\leq r\leq(\log\log x)^2} \exp(O(r))\sum_{d\in S_r}\frac{\rho(d)}{d} \ll \log x. \tag{7.26}$$

It is at this point that we need some algebraic number theory, and specifically the *Landau prime ideal theorem*, via the following lemma:

Proposition 7.3.5. *We have*

$$\sum_{d\leq x}\frac{\rho(d)}{d} \ll \log x. \tag{7.27}$$

Proof. Let k be the number field formed by extending the rationals by adjoining a root α of the irreducible polynomial P. The Landau prime ideal theorem (the generalisation of the prime number theorem to such fields) then tells us (among other things) that the number of prime ideals in k of norm less than x is $x/\log x + O(x/\log^2 x)$. Note that if p is a prime with a simple root $P(n) = 0 \bmod p$ in $\mathbf{Z}/p\mathbf{Z}$, then one can associate a prime ideal in k of norm p defined as $(p, \alpha - n)$. As long as p does not divide the discriminant, one has $\rho(p)$ simple roots; but there are only $O(1)$ primes that divide the discriminant. From this we see that

$$\sum_{p\leq x}\rho(p) \leq \frac{x}{\log x} + O(\frac{x}{\log^2 x}).$$

(One can complement this upper bound with a lower bound, since the ideals whose norms are a power of a (rational) prime rather than a prime have only a negligible contribution to the ideal count, but we will not need the lower bound here). By summation by parts we conclude

$$\sum_{p\leq x}\frac{\rho(p)}{p} \leq \log\log x + O(1)$$

and (7.27) follows by standard multiplicative number theory methods (e.g. bounding $\sum_{d\leq x}\frac{\rho(d)}{d^{1+1/\log x}}$ by computing the Euler product, noting that $\rho(p^j) = \rho(p)$ whenever p does not divide the discriminant of P, thanks to Hensel's lemma). □

This proposition already deals with the bounded r case. For large r we need the following variant:

Proposition 7.3.6. *For any $2 \leq r \leq (\log\log x)^2$, one has*

$$\sum_{d \in S_r} \frac{\rho(d)}{d} \ll r^{-cr} \log x$$

for some absolute constant $c > 0$.

The bound (7.26) then follows as a corollary of this proposition. In fact, one expects the $x^{1/r}$-smoothness in the definition of S_r to induce a gain of about $\frac{1}{r!}$; see [**Gr2008**] for extensive discussion of this and related topics.

Proof. If $d \in S_r$, then we can write $d = p_1 \dots p_j$ for some primes $p_1, \dots, p_j \leq x^{1/r}$. As noted previously, the primes in this product that are less than $x^{1/(\log\log x)^2}$ each contribute at most $x^{1/(\log\log x)^2}$ to this product, and there are at most $O(\log\log x)$ of these primes, so their total contribution is at most $x^{O(1/\log\log x)}$. Since $d \geq x^{1/2}$, we conclude that the primes that are greater than $x^{1/(\log\log x)^2}$ in the factorisation of d must multiply to at least $x^{1/4}$ (say). By definition of S_r, these primes are distinct. By the pigeonhole principle, we can then find $t \geq 1$ such that there are distinct primes $q_1, \dots, q_m$ between $x^{1/2^{t+1}r}$ and $x^{1/2^t r}$ which appear in the prime factorisation of d, where $m := \lfloor \frac{rt}{100} \rfloor$ (say); by definition of S_r, all these primes are distinct and can thus be ordered as $q_1 < \dots < q_m$, and we can write $d = q_1 \dots q_m u$ for some $u \leq x$. As the $\rho(q_j)$ are bounded, we have

$$\rho(d) \ll O(1)^m \rho(u) \ll O(1)^{rt} \rho(u)$$

and so we can upper bound $\sum_{d \in S_r} \frac{\rho(d)}{d}$ by

$$\sum_{t \ll (\log\log x)^2} O(1)^{rt} \sum_{x^{1/2^{t+1}r} \leq q_1 < \dots < q_m \leq x^{1/2^t r}} \frac{1}{q_1 \dots q_m} \sum_{u < x} \rho(u).$$

Using (7.27) and symmetry we can bound this by

$$\sum_{t \ll (\log\log x)^2} O(1)^{rt} \frac{1}{m!} \Big(\sum_{x^{1/2^{t+1}r} \leq q \leq x^{1/2^t r}} \frac{1}{q} \Big)^m \log x.$$

By the prime number theorem (or *Mertens' theorem*) we have

$$\sum_{x^{1/2^{t+1}r} \leq q \leq x^{1/2^t r}} \frac{1}{q} \ll 1.$$

Inserting this bound and summing the series using *Stirling's formula*, one obtains the claim. □

7.4. The Katai-Bourgain-Sarnak-Ziegler asymptotic orthogonality criterion

One of the basic problems in analytic number theory is to estimate sums of the form

$$\sum_{p<x} f(p)$$

as $x \to \infty$, where p ranges over primes and f is some explicit function of interest (e.g. a linear phase function $f(p) = e^{2\pi i \alpha p}$ for some real number α). This is essentially the same task as obtaining estimates on the sum

$$\sum_{n<x} \Lambda(n) f(n)$$

where Λ is the *von Mangoldt function*. If f is bounded, $f(n) = O(1)$, then from the prime number theorem one has the trivial bound

$$\sum_{n<x} \Lambda(n) f(n) = O(x)$$

but often (when f is somehow "oscillatory" in nature) one is seeking the refinement

$$\sum_{n<x} \Lambda(n) f(n) = o(x) \tag{7.28}$$

or equivalently

$$\sum_{p<x} f(p) = o(\frac{x}{\log x}). \tag{7.29}$$

Thanks to identities such as

$$\Lambda(n) = \sum_{d|n} \mu(d) \log(\frac{n}{d}), \tag{7.30}$$

where μ is the *Möbius function*, refinements such as (7.28) are similar in spirit to estimates of the form

$$\sum_{n<x} \mu(n) f(n) = o(x). \tag{7.31}$$

Unfortunately, the connection between (7.28) and (7.31) is not particularly tight; roughly speaking, one needs to improve the bounds in (7.31) (and variants thereof) by about two factors of $\log x$ before one can use identities such as (7.30) to recover (7.28). Still, one generally thinks of (7.28) and (7.31) as being "morally" equivalent, even if they are not formally equivalent.

When f is oscillating in a sufficiently "irrational" way, then one standard way to proceed is the method of Type I and Type II sums, which uses truncated versions of divisor identities such as (7.30) to expand out either (7.28) or (7.31) into linear (Type I) or bilinear sums (Type II) with which

one can exploit the oscillation of f. For instance, Vaughan's identity lets one rewrite the sum in (7.28) as the sum of the Type I sum

$$\sum_{d\leq U} \mu(d)(\sum_{V/d\leq r\leq x/d} (\log r) f(rd)),$$

the Type I sum

$$-\sum_{d\leq UV} a(d) \sum_{V/d\leq r\leq x/d} f(rd),$$

the Type II sum

$$-\sum_{V\leq d\leq x/U} \sum_{U<m\leq x/V} \Lambda(d) b(m) f(dm),$$

and the error term $\sum_{d\leq V} \Lambda(n) f(n)$, whenever $1 \leq U, V \leq x$ are parameters, and a, b are the sequences

$$a(d) := \sum_{e\leq U, f\leq V: ef=d} \Lambda(d)\mu(e)$$

and

$$b(m) := \sum_{d|m: d\leq U} \mu(d).$$

Similarly one can express (7.31) as the Type I sum

$$-\sum_{d\leq UV} c(d) \sum_{UV/d\leq r\leq x/d} f(rd),$$

the Type II sum

$$-\sum_{V<d\leq x/U} \sum_{U<m\leq x/d} \mu(m) b(d) f(dm)$$

and the error term $\sum_{d\leq UV} \mu(n) f(N)$, whenever $1 \leq U, V \leq x$ with $UV \leq x$, and c is the sequence

$$c(d) := \sum_{e\leq U, f\leq V: ef=d} \mu(d)\mu(e).$$

After eliminating troublesome sequences such as $a(), b(), c()$ via Cauchy-Schwarz or the triangle inequality, one is then faced with the task of estimating Type I sums such as

$$\sum_{r\leq y} f(rd)$$

or Type II sums such as

$$\sum_{r\leq y} f(rd)\overline{f(rd')}$$

for various $y, d, d' \geq 1$. Here, the trivial bound is $O(y)$, but due to a number of logarithmic inefficiencies in the above method, one has to obtain bounds

that are more like $O(\frac{y}{\log^C y})$ for some constant C (e.g. $C = 5$) in order to end up with an asymptotic such as (7.28) or (7.31).

However, in a recent paper [**BoSaZi2011**] of Bourgain, Sarnak, and Ziegler, it was observed that as long as one is only seeking the Mobius orthogonality (7.31) rather than the von Mangoldt orthogonality (7.28), one can avoid losing any logarithmic factors, and rely purely on qualitative equidistribution properties of f. A special case of their orthogonality criterion (which had been discovered previously by Kátai [**Ka1986**]) is as follows:

Proposition 7.4.1 (Orthogonality criterion)**.** *Let $f : \mathbf{N} \to \mathbf{C}$ be a bounded function such that*

$$\sum_{n \leq x} f(pn)\overline{f(qn)} = o(x) \tag{7.32}$$

for any distinct primes p, q (where the decay rate of the error term $o(x)$ may depend on p and q). Then

$$\sum_{n \leq x} \mu(n) f(n) = o(x). \tag{7.33}$$

Actually, the Bourgain-Sarnak-Ziegler paper establishes a more quantitative version of this proposition, in which μ can be replaced by an arbitrary bounded multiplicative function, but we will content ourselves with the above weaker special case. This criterion can be viewed as a multiplicative variant of the classical van der Corput lemma, which in our notation asserts that $\sum_{n \leq x} f(n) = o(x)$ if one has $\sum_{n \leq x} f(n+h)\overline{f(n)} = o(x)$ for each fixed non-zero h.

As a sample application, Proposition 7.4.1 easily gives a proof of the asymptotic

$$\sum_{n \leq x} \mu(n) e^{2\pi i \alpha n} = o(x)$$

for any irrational α. (For rational α, this is a little trickier, as it is basically equivalent to the prime number theorem in arithmetic progressions.) In [**BoSaZi2011**] this criterion is also applied to nilsequences (obtaining a quick proof of a qualitative version of a result in [**GrTa2012**]) and to horocycle flows (for which no Möbius orthogonality result was previously known).

Informally, the connection between (7.32) and (7.33) comes from the multiplicative nature of the Möbius function. If (7.33) failed, then $\mu(n)$ exhibits strong correlation with $f(n)$; by change of variables, we then expect $\mu(pn)$ to correlate with $f(pn)$ and $\mu(pm)$ to correlate with $f(qn)$, for "typical" p, q at least. On the other hand, since μ is multiplicative, $\mu(pn)$ exhibits

strong correlation with $\mu(qn)$. Putting all this together (and pretending correlation is transitive), this would give the claim (in the contrapositive). Of course, correlation is not quite transitive, but it turns out that one can use the Cauchy-Schwarz inequality as a substitute for transitivity of correlation in this case.

We will give a proof of Proposition 7.4.1 shortly. The main idea is to exploit the following observation: if P is a "large" but finite set of primes (in the sense that the sum $A := \sum_{p\in P} \frac{1}{p}$ is large), then for a typical large number n (much larger than the elements of P), the number of primes in P that divide n is pretty close to $A = \sum_{p\in P} \frac{1}{p}$:

$$\sum_{p\in P: p|n} 1 \approx A. \tag{7.34}$$

A more precise formalisation of this heuristic is provided by the *Turan-Kubilius inequality*, which is proven by a simple application of the *second moment method*.

In particular, one can sum (7.34) against $\mu(n)f(n)$ and obtain an approximation

$$\sum_{n\leq x} \mu(n)f(n) \approx \frac{1}{A} \sum_{p\in P} \sum_{n\leq x: p|n} \mu(n)f(n)$$

that approximates a sum of $\mu(n)f(n)$ by a bunch of sparser sums of $\mu(n)f(n)$. Since

$$x = \frac{1}{A} \sum_{p\in P} \frac{x}{p},$$

we see (heuristically, at least) that in order to establish (7.31), it would suffice to establish the sparser estimates

$$\sum_{n\leq x: p|n} \mu(n)f(n) = o(\frac{x}{p})$$

for all $p \in P$ (or at least for "most" $p \in P$).

Now we make the change of variables $n = pm$. As the Möbius function is multiplicative, we usually have $\mu(n) = \mu(p)\mu(m) = -\mu(m)$. (There is an exception when n is divisible by p^2, but this will be a rare event and we will be able to ignore it). So it should suffice to show that

$$\sum_{m\leq x/p} \mu(m)f(pm) = o(x/p)$$

for most $p \in P$. However, by the hypothesis (7.32), the sequences $m \mapsto f(pm)$ are asymptotically orthogonal as p varies, and this claim will then follow from a Cauchy-Schwarz argument.

7.4.1. Rigorous proof. We will need a slowly growing function $H = H(x)$ of x, with $H(x) \to \infty$ as $x \to \infty$, to be chosen later. As the sum of reciprocals of primes diverges, we see that

$$\sum_{p<H} \frac{1}{p} \to \infty$$

as $x \to \infty$. It will also be convenient to eliminate small primes. Note that we may find an even slower growing function $W = W(x)$ of x, with $W(x) \to \infty$ as $x \to \infty$, such that

$$\sum_{W \leq p<H} \frac{1}{p} \to \infty.$$

Although it is not terribly important, we will take W and H to be powers of two. Thus, if we set P to be all the primes between W and H, the quantity

$$A := \sum_{p \in P} \frac{1}{p}$$

goes to infinity as $x \to \infty$.

Lemma 7.4.2 (Turan-Kubilius inequality)**.** *One has*

$$\sum_{n \leq x} |\sum_{p \in P : p|n} 1 - A|^2 \ll Ax. \tag{7.35}$$

Proof. We have

$$\sum_{n \leq x} \sum_{p \in P : p|n} 1 = \sum_{p \in P} \sum_{n \leq x : p|n} 1.$$

On the other hand, we have

$$\sum_{n \leq x : p|n} 1 = \frac{x}{p} + O(1)$$

and thus (if H is sufficiently slowly growing)

$$\sum_{n \leq x} \sum_{p \in P : p|n} 1 = xA + O(x).$$

Similarly, we have

$$\sum_{n \leq x} (\sum_{p \in P : p|n} 1)^2 = \sum_{p,q \in P} \sum_{n \leq x : p|n, q|n} 1.$$

The expression $\sum_{n \leq x : p|n,q|n} 1$ is equal to $\frac{x}{p} + O(1)$ when $q = p$, and $\frac{x}{pq}$ when $p \neq q$. A brief calculation then shows that

$$\sum_{n \leq x} (\sum_{p \in P : p|n} 1)^2 = xA^2 + O(Ax)$$

if H is sufficiently slowly growing. Inserting these bounds into (7.35), the claim follows. □

From (7.35) and the Cauchy-Schwarz inequality, one has

$$\sum_{n\leq x}(\sum_{p\in P:p|n} 1 - A)\mu(n)f(n) = O(A^{1/2}x)$$

which we rearrange as

$$\sum_{n\leq x}\mu(n)f(n) = \frac{1}{A}\sum_{p\in P}\sum_{n\leq x:p|n}\mu(n)f(n) + O(A^{-1/2}x).$$

Since A goes to infinity, the $O(A^{-1/2}x)$ term is $o(x)$, so it now suffices to show that

$$\sum_{p\in P}\sum_{n\leq x:p|n}\mu(n)f(n) = o(Ax).$$

Write $n = pm$. Then we have $\mu(n)f(n) = -\mu(m)f(pm)$ for all but $O(x/p^2)$ values of n (if H is sufficiently slowly growing). The exceptional values contribute at most

$$\sum_{p\in P}\frac{x}{p^2} = \sum_{p\in P}\frac{x}{Wp} = O(Ax/W) = o(Ax)$$

which is acceptable. Thus it suffices to show that

$$\sum_{p\in P}\sum_{m\leq x/p}\mu(m)f(pm) = o(Ax).$$

Partitioning into dyadic blocks, it suffices to show that

$$\sum_{p\in P_k}\sum_{m\leq x/p}\mu(m)f(pm) = o(|P_k|x/2^k)$$

uniformly for $W \leq 2^k < H$, where P_k are the primes between 2^k and 2^{k+1}.

Fix k. The left-hand side can be rewritten as

$$\sum_{m\leq x/2^k}\mu(m)\sum_{p\in P_k} f(pm)1_{m\leq x/p}$$

so by the Cauchy-Schwarz inequality it suffices to show that

$$\sum_{m\leq x/2^k}|\sum_{p\in P_k} f(pm)1_{m\leq x/p}|^2 = o(|P_k|^2 x/2^k).$$

We can rearrange the left-hand side as

$$\sum_{p,q\in P_k}\sum_{m\leq \min(x/p,x/q)} f(pm)\overline{f(qm)}.$$

Now if H is sufficiently slowly growing as a function of x, we see from (7.32) that for all distinct $p, q \leq H$, we have

$$\sum_{m\leq \min(x/p,x/q)} f(pm)\overline{f(qm)} = o(x/2^k)$$

uniformly in p, q; meanwhile, for $p = q$, we have the crude bound

$$\sum_{m \leq \min(x/p, x/q)} f(pm)\overline{f(qm)} = O(x/2^k).$$

The claim follows (noting from the prime number theorem that $|P_k| = o(|P_k|^2)$).

7.4.2. From Möbius to von Mangoldt? It would be great if one could pass from the Möbius asymptotic orthogonality (7.31) to the von Mangoldt asymptotic orthgonality (7.28) (or equivalently, to (7.29)), as this would give some new information about the distribution of primes. Unfortunately, it seems that some additional input is needed to do so. Here is a simple example of a conditional implication that requires an additional input, namely some quantitative control on "Type I" sums:

Proposition 7.4.3. *Let $f : \mathbf{N} \to \mathbf{C}$ be a bounded function such that*

$$\sum_{n \leq x} \mu(n) f(dn) = o(x) \tag{7.36}$$

for each fixed $d \geq 1$ (with the decay rate allowed to depend on d). Suppose also that one has the Type I bound

$$\sum_{1 \leq m \leq M} \sup_{y \leq x} |\sum_{n \leq y} f(mn)| \ll \frac{Mx}{\log^{2+\varepsilon} x} \tag{7.37}$$

for all $M, x \geq 2$ and some absolute constant $\varepsilon > 0$, where the implied constant is independent of both M and x. Then one has

$$\sum_{n \leq x} \Lambda(n) f(n) = o(x) \tag{7.38}$$

and thus (by discarding the prime powers and summing by parts)

$$\sum_{p \in x} f(p) = o(\frac{x}{\log x}).$$

Proof. We use the Dirichlet hyperbola method. Using (7.30), one can write the left-hand side of (7.38) as

$$\sum_{dm \leq x} \mu(m)(\log d) f(dm).$$

We let $D = D(x)$ be a slowly growing function of x to be chosen later, and split this sum as

$$\sum_{d \leq D} \log d \sum_{m \leq x/d} \mu(m) f(dm) + \sum_{m < x/D} \mu(m) \sum_{D < d \leq x/m} (\log d) f(dm). \tag{7.39}$$

If D is sufficiently slowly growing, then by (7.36) one has $\sum_{m \leq x/d} \mu(m) f(dm) = o(x)$ uniformly for all $d \leq D$. If D is sufficiently slowly growing, this implies that the first term in (7.39) is also $o(x)$. As for the second term, we dyadically decompose it and bound it in absolute value by

$$\sum_{2^k < x/D} \sum_{2^k \leq m < 2^{k+1}} |\sum_{D < d \leq x/m} (\log d) f(dm)|. \tag{7.40}$$

By summation by parts, we can bound

$$|\sum_{D < d \leq x/m} (\log d) f(dm)| \ll \log(x/2^k) \sup_{y \leq x/2^k} |\sum_{n \leq y} f(mn)|$$

and so by (7.37), we can bound (7.40) by

$$\ll \sum_{2^k < x/D} \frac{x}{\log^{1+\varepsilon}(x/2^k)}.$$

This sum evaluates to $O(x/D^\varepsilon)$, and the claim follows since D goes to infinity. □

Note that the trivial bound on (7.37) is Mx, so one needs to gain about two logarithmic factors over the trivial bound in order to use the above proposition. The presence of the supremum is annoying, but it can be removed by a modification of the argument if one improves the bound by an additional logarithm by a variety of methods (e.g. completion of sums), or by smoothing out the constraint $n \leq x$. However, I do not know of a way to remove the need to improve the trivial bound by two logarithmic factors.

Chapter 8

Geometry

8.1. A geometric proof of the impossibility of angle trisection by straightedge and compass

One of the most well known problems from ancient Greek mathematics was that of trisecting an angle by straightedge and compass, which was eventually proven impossible in [**Wa1836**], using methods from Galois theory.

Formally, one can set up the problem as follows. Define a *configuration* to be a finite collection $\mathcal{C}$ of points, lines, and circles in the Euclidean plane. Define a *construction step* to be one of the following operations to enlarge the collection $\mathcal{C}$:

- (Straightedge) Given two distinct points A, B in $\mathcal{C}$, form the line $\overline{AB}$ that connects A and B, and add it to $\mathcal{C}$.
- (Compass) Given two distinct points A, B in $\mathcal{C}$, and given a third point O in $\mathcal{C}$ (which may or may not equal A or B), form the circle with centre O and radius equal to the length $|AB|$ of the line segment joining A and B, and add it to $\mathcal{C}$.
- (Intersection) Given two distinct curves γ, γ' in $\mathcal{C}$ (thus γ is either a line or a circle in $\mathcal{C}$, and similarly for γ'), select a point P that is common to both γ and γ' (there are at most two such points), and add it to $\mathcal{C}$.

We say that a point, line, or circle is *constructible by straightedge and compass* from a configuration $\mathcal{C}$ if it can be obtained from $\mathcal{C}$ after applying a finite number of construction steps.

Problem 8.1.1 (Angle trisection)**.** Let A, B, C be distinct points in the plane. Is it always possible to construct by straightedge and compass from

A, B, C a line ℓ through A that *trisects* the angle $\angle BAC$, in the sense that the angle between ℓ and BA is one third of the angle of $\angle BAC$?

Thanks to Wantzel's result [**Wa1836**], the answer to this problem is known to be "no" in general; a *generic* angle $\angle BAC$ cannot be trisected by straightedge and compass. (On the other hand, some *special* angles can certainly be trisected by straightedge and compass, such as a right angle. Also, one can certainly trisect generic angles if other methods than straightedge and compass are permitted.)

The impossibility of angle trisection stands in sharp contrast to the easy construction of angle *bisection* via straightedge and compass, which we briefly review as follows:

(1) Start with three points A, B, C.

(2) Form the circle c_0 with centre A and radius AB, and intersect it with the line $\overline{AC}$. Let D be the point in this intersection that lies on the same side of A as C. (D may well be equal to C.)

(3) Form the circle c_1 with centre B and radius AB, and the circle c_2 with centre D and radius AB. Let E be the point of intersection of c_1 and c_2 that is not A.

(4) The line $\ell := \overline{AE}$ will then bisect the angle $\angle BAC$.

See Figure 1. The key difference between angle trisection and angle bisection ultimately boils down to the following trivial number-theoretic fact:

Lemma 8.1.2. *There is no power of* 2 *that is evenly divisible by* 3*.*

Proof. Obvious by modular arithmetic, by induction, or by the fundamental theorem of arithmetic. □

In contrast, there are of course plenty of powers of 2 that are evenly divisible by 2, and this is ultimately why angle bisection is easy while angle trisection is hard.

The standard way in which Lemma 8.1.2 is used to demonstrate the impossibility of angle trisection is via *Galois theory*. The implication is quite short if one knows this theory, but quite opaque otherwise. We briefly sketch the proof of this implication here, though we will not need it in the rest of the discussion. Firstly, Lemma 8.1.2 implies the following fact about *field extensions*.

Corollary 8.1.3. *Let F be a field, and let E be an extension of F that can be constructed out of F by a finite sequence of* quadratic extensions*. Then E does not contain any cubic extensions K of F.*

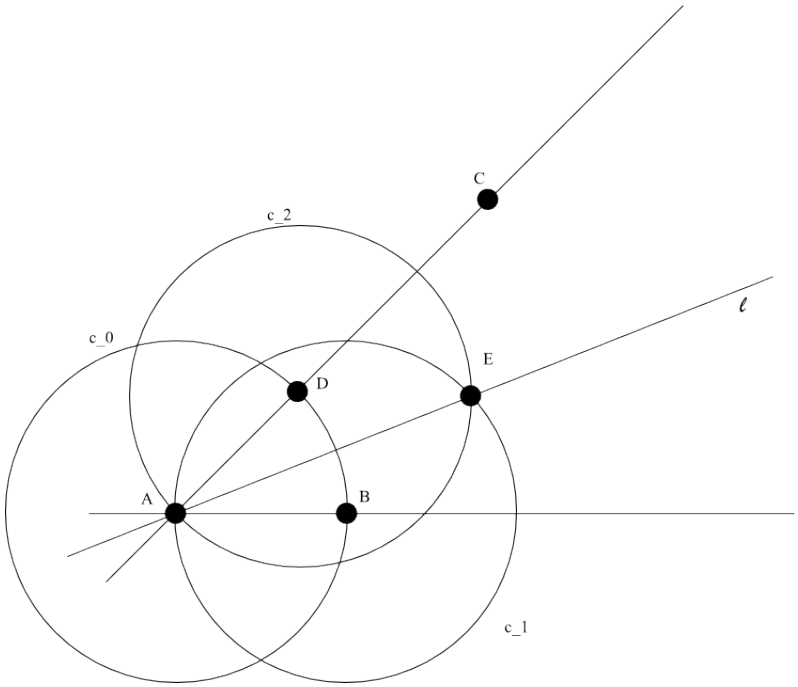

Figure 1. Angle bisection.

Proof. If E contained a cubic extension K of F, then the dimension of E over F would be a multiple of three. On the other hand, if E is obtained from F by a tower of quadratic extensions, then the dimension of E over F is a power of two. The claim then follows from Lemma 8.1.2. □

To conclude the proof, one then notes that any point, line, or circle that can be constructed from a configuration $\mathcal{C}$ is definable in a field obtained from the coefficients of all the objects in $\mathcal{C}$ after taking a finite number of quadratic extensions, whereas a trisection of an angle $\angle ABC$ will generically only be definable in a cubic extension of the field generated by the coordinates of A, B, C.

The Galois theory method also allows one to obtain many other impossibility results of this type, most famously the *Abel-Ruffini theorem* on the insolvability of the quintic equation by radicals. For this reason (and also because of the many applications of Galois theory to number theory and other branches of mathematics), the Galois theory argument is the "right" way to prove the impossibility of angle trisection within the broader framework of modern mathematics. However, this argument has the drawback that it requires one to first understand Galois theory (or at least field theory), which is usually not presented until an advanced undergraduate algebra or number

theory course, whilst the angle trisection problem requires only high-school level mathematics to formulate. Even if one is allowed to "cheat" and sweep several technicalities under the rug, one still needs to possess a fair amount of solid intuition about advanced algebra in order to appreciate the proof. (This was undoubtedly one reason why, even after Wantzel's impossibility result was published, a large amount of effort was still expended by amateur mathematicians to try to trisect a general angle.)

In this section, I would therefore like to present a different proof (or perhaps more accurately, a disguised version of the standard proof) of the impossibility of angle trisection by straightedge and compass, that avoids explicit mention of Galois theory (though it is never far beneath the surface). With "cheats", the proof is actually quite simple and geometric (except for Lemma 8.1.2, which is still used at a crucial juncture), based on the basic geometric concept of *monodromy*; unfortunately, some technical work is needed however to remove these cheats.

To describe the intuitive idea of the proof, let us return to the angle bisection construction, that takes a triple A, B, C of points as input and returns a bisecting line ℓ as output. We iterate the construction to create a quartisecting line m, via the following sequence of steps that extend the original bisection construction:

(1) Start with three points A, B, C.

(2) Form the circle c_0 with centre A and radius AB, and intersect it with the line $\overline{AC}$. Let D be the point in this intersection that lies on the same side of A as C. (D may well be equal to C.)

(3) Form the circle c_1 with centre B and radius AB, and the circle c_2 with centre D and radius AB. Let E be the point of intersection of c_1 and c_2 that is not A.

(4) Let F be the point on the line $\ell := \overline{AE}$ which lies on c_0, and is on the same side of A as E.

(5) Form the circle c_3 with centre F and radius AB. Let G be the point of intersection of c_1 and c_3 that is not A.

(6) The line $m := \overline{AG}$ will then quartisect the angle $\angle BAC$.

See Figure 2. Let us fix the points A and B, but not C, and view m (as well as intermediate objects such as D, c_2, E, ℓ, F, c_3, G) as a function of C.

Let us now do the following: we begin rotating C counterclockwise around A, which drags around the other objects D, c_2, E, ℓ, F, c_3, G that were constructed by C accordingly. For instance, here is an early stage of

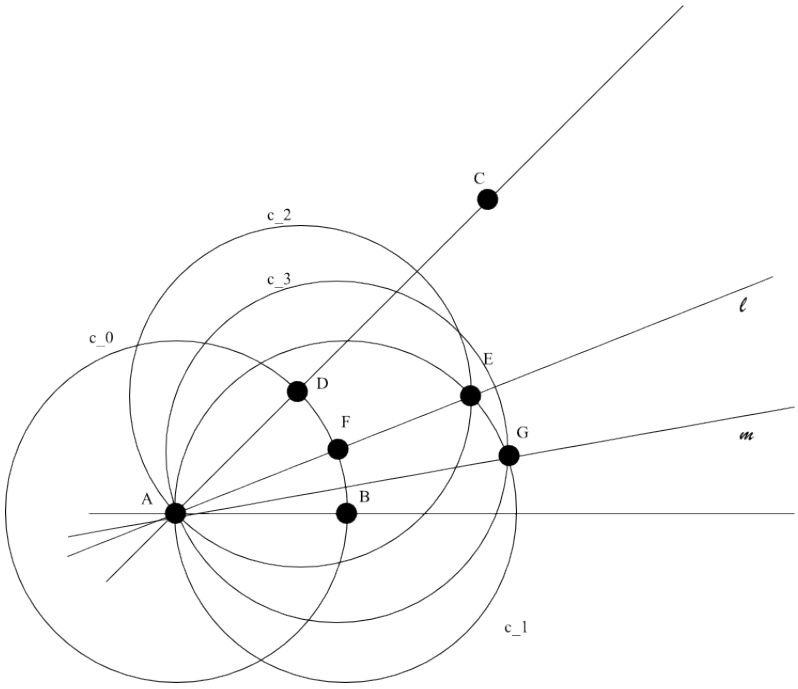

Figure 2. Angle quadrisection.

this rotation process, when the angle $\angle BAC$ has become obtuse; see Figure 3.

Now for the slightly tricky bit. We are going to keep rotating C beyond a half-rotation of 180°, so that $\angle BAC$ now becomes a *reflex angle*. At this point, a singularity occurs; the point E collides into A, and so there is an instant in which the line $\ell = \overline{AE}$ is not well-defined. However, this turns out to be a *removable singularity* (and the easiest way to demonstrate this will be to tap the power of complex analysis, as complex numbers can easily route around such a singularity), and we can blast through it to the other side; see Figure 4.

Note that we have now deviated from the original construction in that F and E are no longer on the same side of A; we are thus now working in a *continuation* of that construction rather than with the construction itself. Nevertheless, we can still work with this continuation (much as, say, one works with *analytic continuations* of infinite series such as $\sum_{n=1}^{\infty} \frac{1}{n^s}$ beyond their original domain of definition).

We now keep rotating C around A. In Figure 5, $\angle BAC$ is approaching a full rotation of 360°.

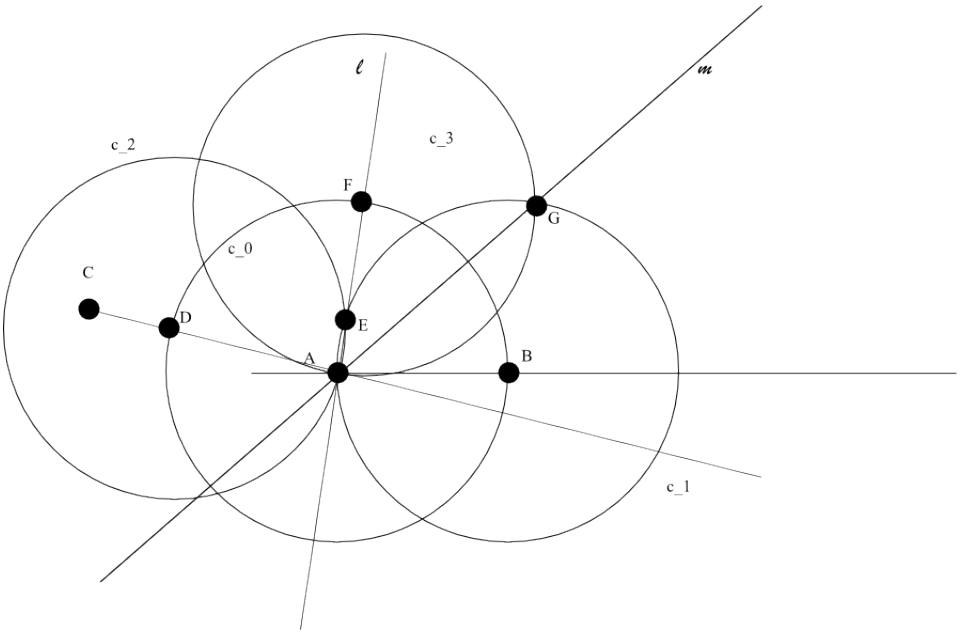

Figure 3. $\angle BAC$ becomes obtuse.

When $\angle BAC$ reaches a full rotation, a different singularity occurs: c_1 and c_2 coincide. Nevertheless, this is also a removable singularity, and we blast through to beyond a full rotation; see Figure 6.

And now C is back where it started, as are D, c_2, E, and ℓ... but the point F has moved, from one intersection point of $\ell \cap c_3$ to the other. As a consequence, c_3, G, and m have also changed, with m being at right angles to where it was before. (In the jargon of modern mathematics, the quadrisection construction has a non-trivial *monodromy*.)

But nothing stops us from rotating C some more. If we continue this procedure, we see that after two full rotations of C around A, all points, lines, and circles constructed from A, B, C have returned to their original positions. Because of this, we shall say that the quadrisection construction described above is *periodic with period* 2.

Similarly, if one performs an octisection of the angle $\angle BAC$ by bisecting the quadrisection, one can verify that this octisection is periodic with period 4; it takes four full rotations of C around A before the configuration returns to where it started. More generally, one can show

Proposition 8.1.4. *Any construction of straightedge and compass from the points A, B, C is periodic with period equal to a power of* 2.

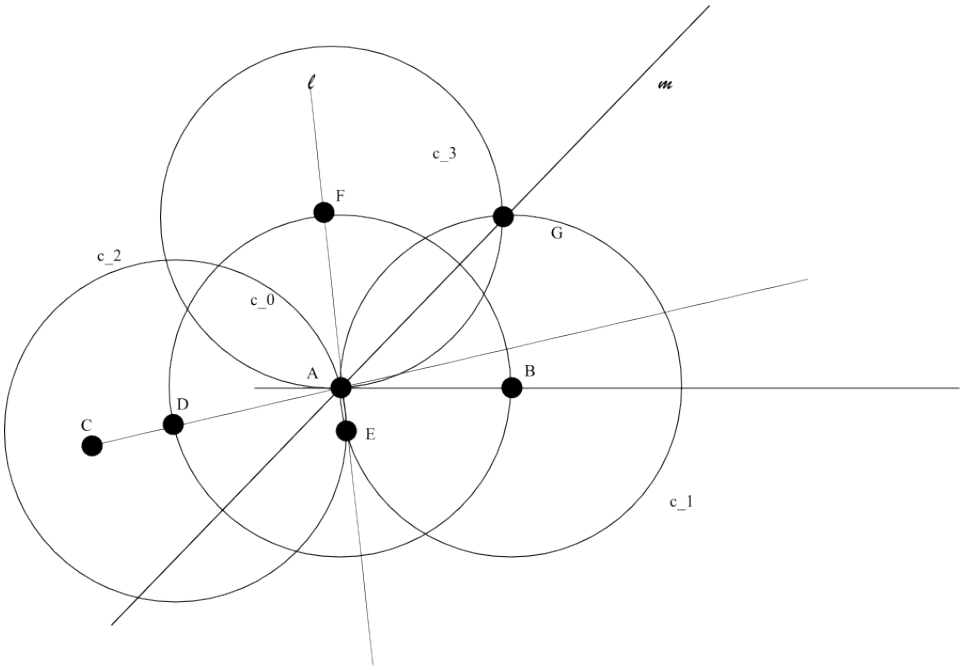

Figure 4. Beyond the removable singularity.

The reason for this, ultimately, is because any two circles or lines will intersect each other in at most two points, and so at each step of a straightedge-and-compass construction there is an ambiguity of at most $2! = 2$. Each rotation of C around A can potentially flip one of these points to the other, but then if one rotates again, the point returns to its original position, and then one can analyse the next point in the construction in the same fashion until one obtains the proposition.

But now consider a putative trisection operation, that starts with an arbitrary angle $\angle BAC$ and somehow uses some sequence of straightedge and compass constructions to end up with a trisecting line ℓ: see Figure 7.

What is the period of this construction? If we continuously rotate C around A, we observe that a full rotations of C only causes the trisecting line ℓ to rotate by a third of a full rotation (i.e. by $120°$): see Figure 8.

Because of this, we see that the period of any construction that contains ℓ must be a multiple of 3. But this contradicts Proposition 8.1.4 and Lemma 8.1.2.

We will now make the above proof rigorous. Unfortunately, in doing so, one has to leave the world of high-school mathematics, as one needs a little bit of algebraic geometry and complex analysis to resolve the issues with singularities that we saw in the above sketch. Still, I feel that at an

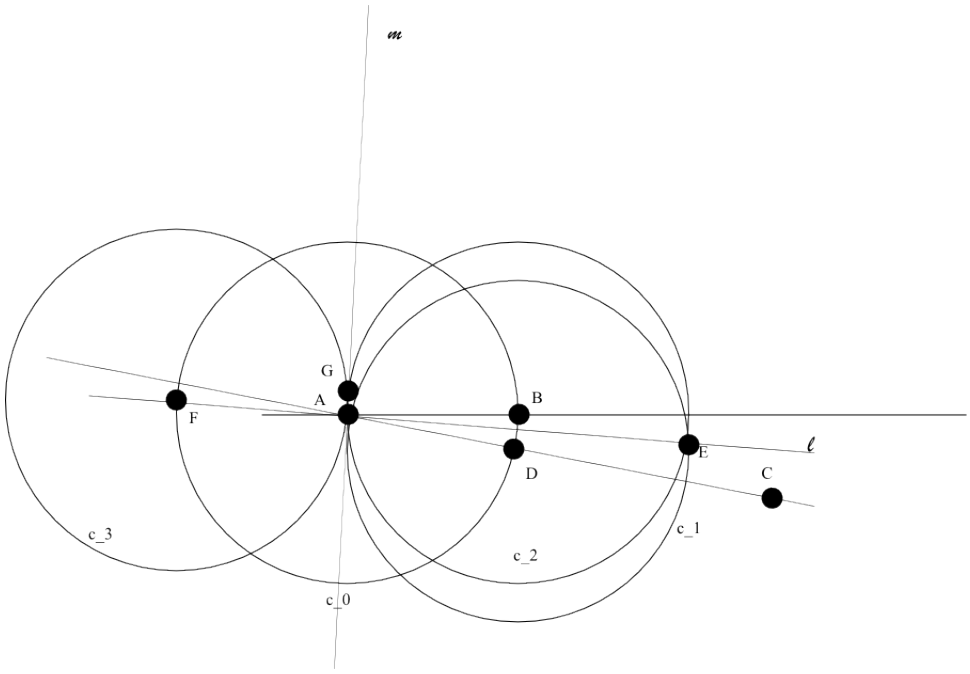

Figure 5. $\angle BAC$ is almost full.

intuitive level at least, this argument is more geometric and accessible than the Galois-theoretic argument (though anyone familiar with Galois theory will note that there is really not that much difference between the proofs, ultimately, as one has simply replaced the Galois group with a closely related monodromy group instead).

8.1.1. Details. We will assume for sake of contradiction that for every triple A, B, C of distinct points, we can find a construction by straightedge and compass that trisects the angle $\angle BAC$, and eventually deduce a contradiction out of this.

We remark that we do not initially assume any uniformity in this construction; for instance, it could be possible that the trisection procedure for obtuse angles is completely different from that of acute angles, using a totally different set of constructions, while some exceptional angles (e.g. right angles or degenerate angles) might use yet another construction. We will address these issues later.

The first step is to get rid of some possible degeneracies in one's construction. At present, nothing in our definition of a construction prevents us from adding a point, line, or circle to the construction that was already present in the existing collection $\mathcal{C}$ of points, lines, and circles. However, it is

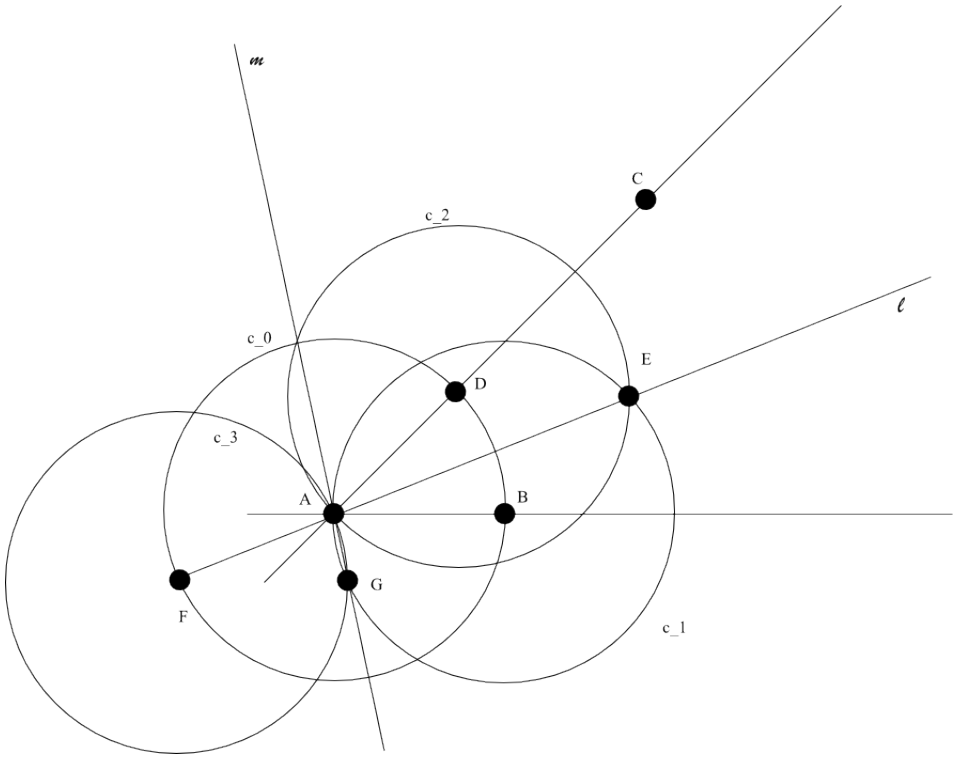

Figure 6. $\angle BAC$ is beyond full.

clear that any such step in the construction is redundant, and can be omitted. Thus, we may assume without loss of generality that for each A, B, C, the construction used to trisect the angle contains no such redundant steps. (This may make the construction even less uniform than it was previously, but we will address this issue later.)

Another form of degeneracy that we will need to eliminate for technical reasons is that of *tangency*. At present, we allow in our construction the ability to take two tangent circles, or a circle and a tangent line, and add the tangent point to the collection (if it was not already present in the construction). This would ordinarily be a harmless thing to do, but it complicates our strategy of perturbing the configuration, so we now act to eliminate it. Suppose first that one had two circles c_1, c_2 already constructed in the configuration $\mathcal{C}$ and tangent to each other, and one wanted to add the tangent point T to the configuration. But note that in order to have added c_1 and c_2 to $\mathcal{C}$, one must previously have added the centres A_1 and A_2 of these circles to $\mathcal{C}$ also. One can then add T to $\mathcal{C}$ by intersecting the line $\overline{A_1 A_2}$ with c_1 and picking the point that lies on c_2; this way, one does not need to intersect two tangent curves together: see Figure 9.

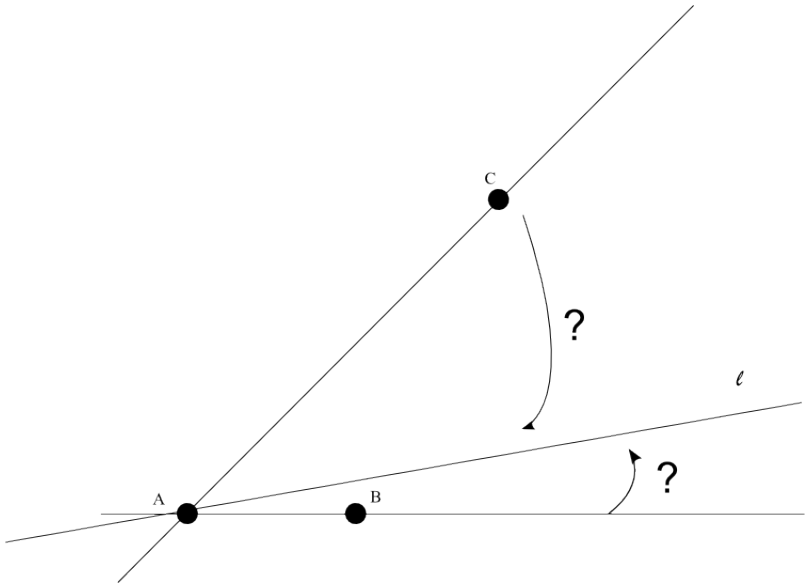

Figure 7. A putative trisection operation.

Similarly, suppose that we already had a circle c and a tangent line ℓ already constructed in the configuration, but with the tangent point T absent. The centre A of c, and at least two points B, C on ℓ, must previously have also been constructed in order to have c and ℓ present; note that B, C are not equal to T by hypothesis. One can then obtain T by dropping a perpendicular from A to ℓ by the usual construction (i.e. drawing a circle centred at A with radius $|AB|$ to hit ℓ again at D, then drawing circles from B and D with the same radius $|AB|$ to meet at a point E distinct from A, then intersecting AE with ℓ to obtain T), thus[1] avoiding tangencies again; see Figure 10.

As a consequence of these reductions, we may now assume that our construction is *nondegenerate* in the sense that

- Any point, line, or circle added at a step in the construction, does not previously appear in that construction.
- Whenever one intersects two circles in a construction together to add another point to the construction, the circles are non-tangent (and thus meet in exactly two points).
- Whenever one intersects a circle and a line in a construction together to add another point to the construction, the circle and line are non-tangent (and thus meet in exactly two points).

[1]This construction may happen to use lines or circles that had already appeared in the construction, but in those cases one can simply skip those steps.

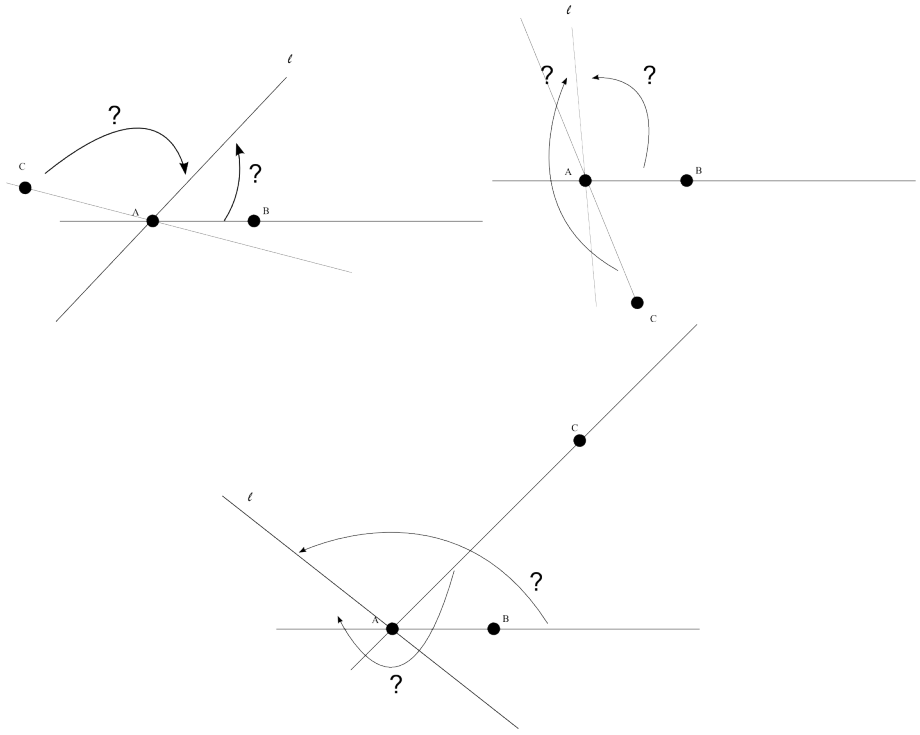

Figure 8. Rotating the trisection.

The reason why we restrict attention to nondegenerate constructions is that they are *stable with respect to perturbations*. Note for instance that if one has two circles c_1, c_2 that intersect in two different points, and one of them is labeled P, then we may perturb c_1 and c_2 by a small amount, and still have an intersection point close to P (with the other intersection point far away from P). Thus, P is locally a continuous function of c_1 and c_2. Similarly if one forms the intersection of a circle and a secant (a line which intersects non-tangentially). In a similar vein, given two points A and B that are distinct, the line between them $\overline{AB}$ varies continuously with A and B as long as one does not move A and B so far that they collide; and given two lines ℓ_1 and ℓ_2 that intersect at a point C (and in particular are non-parallel), then C also depends continuously on ℓ_1 and ℓ_2. Thus, in a nondegenerate construction starting from the original three points A, B, C, every point, line, or circle created by the construction can be viewed as a continuous function of A, B, C, as long as one only works in a sufficiently small neighbourhood of the original configuration (A, B, C). In particular, the final line ℓ varies continuously in this fashion. Note however that the trisection property may be lost by this perturbation; just because ℓ happens to trisect $\angle BAC$ when A, B, C are in the original positions, this does not necessarily imply that

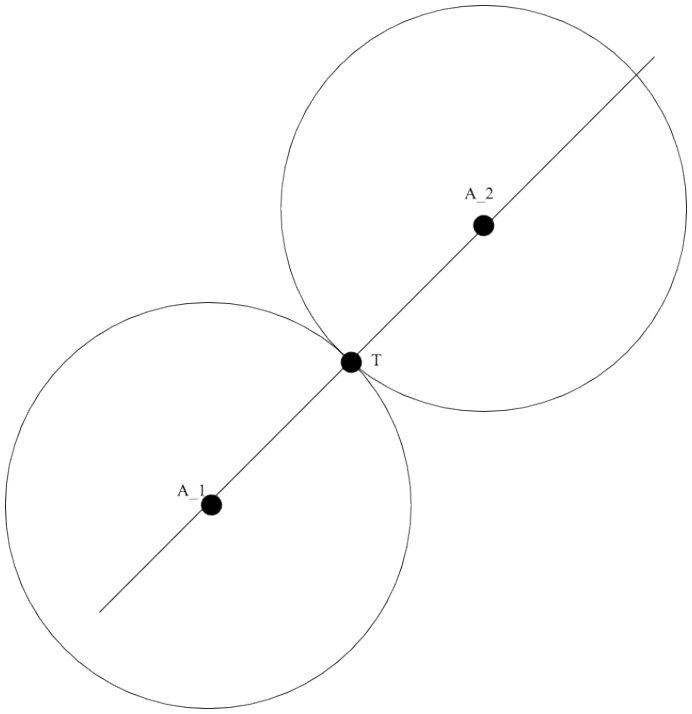

Figure 9. Eliminating tangency.

after one perturbs A, B, C, that the resulting perturbed line ℓ still trisects the angle. (For instance, there are a number of ways to trisect a right angle (e.g. by bisecting an angle of an equilateral triangle), but if one perturbs the angle to be slightly acute or slightly obtuse, the line created by this procedure would not be expected to continue to trisect that angle.)

The next step is to allow *analytic geometry* (and thence *algebraic geometry*) to enter the picture, by using *Cartesian coordinates*. We may identify the Euclidean plane with the analytic plane $\mathbf{R}^2 := \{(x, y) : x, y \in \mathbf{R}\}$; we may also normalise A, B to be the points $A = (0, 0)$, $B = (1, 0)$ by this identification. We will also restrict C to lie on the unit circle $S^1 := \{(x, y) \in \mathbf{R}^2 : x^2 + y^2 = 1\}$, so that there is now just one degree of freedom in the configuration (A, B, C). One can describe a line in $\mathbf{R}^2$ by an equation of the form

$$\{(x, y) \in \mathbf{R}^2 : ax + by + c = 0\}$$

(with a, b not both zero), and describe a circle in $\mathbf{R}^2$ by an equation of the form

$$\{(x, y) \in \mathbf{R}^2 : (x - x_0)^2 + (y - y_0)^2 = r^2\}$$

with r non-zero. There is some non-uniqueness in these representations: for the line, one can multiply a, b, c by the same constant without altering the

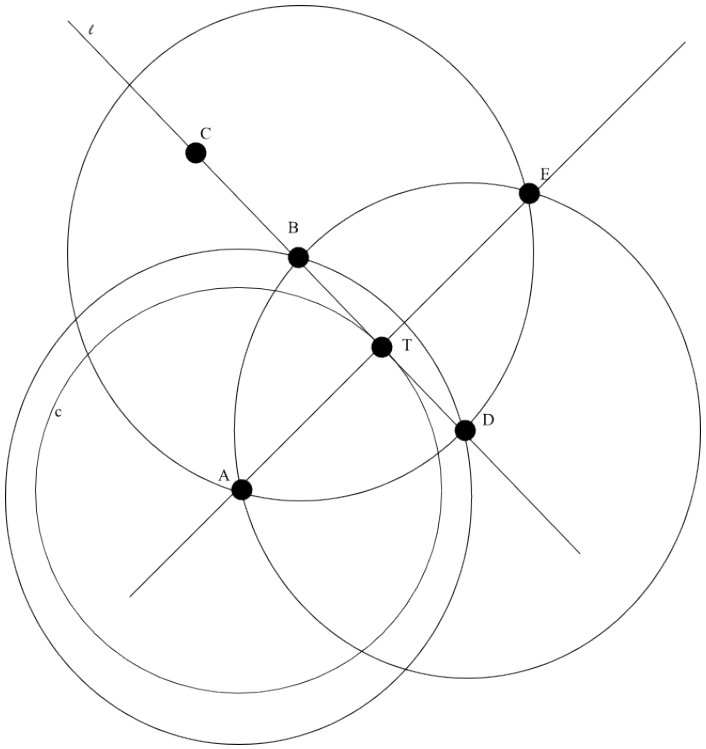

Figure 10. Another tangency elimination.

line, and for the circle, one can replace r by $-r$. However, this will not be a serious concern for us. Note that any two distinct points $P = (x_1, y_1)$, $Q = (x_2, y_2)$ determine a line

$$\{(x, y) \in \mathbf{R}^2 : xy_1 - xy_2 - yx_1 + yx_2 + x_1y_2 - x_2y_1 = 0\}$$

and given three points $O = (x_0, y_0)$, $A = (x_1, y_1)$, $B = (x_2, y_2)$, one can form a circle

$$\{(x, y) \in \mathbf{R}^2 : (x - x_0)^2 + (y - y_0)^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2\}$$

with centre O and radius $|AB|$. Given two distinct non-parallel lines

$$\ell = \{(x, y) \in \mathbf{R}^2 : ax + by + c = 0\}$$

and

$$\ell' = \{(x, y) \in \mathbf{R}^2 : a'x + b'y + c' = 0\},$$

their unique intersection point is given as

$$(\frac{bc' - b'c}{ab' - ba'}, \frac{a'c - c'a}{ab' - ba'});$$

similarly, given two circles

$$c_1 = \{(x,y) \in \mathbf{R}^2 : (x-x_1)^2 + (y-y_1)^2 = r_1^2\}$$

and

$$c_2 = \{(x,y) \in \mathbf{R}^2 : (x-x_2)^2 + (y-y_2)^2 = r_2^2\},$$

their points of intersection (if they exist in $\mathbf{R}^2$) are given as

$$(x_1,y_1) + t(x_2-x_1, y_2-y_1) \pm (\frac{r^2 - t^2 d^2}{d^2})^{1/2}(y_1-y_2, x_2-x_1) \tag{8.1}$$

where

$$t := \frac{1}{2} - \frac{r_2^2 - r_1^2}{2d^2}$$

and

$$d^2 := (x_1-x_2)^2 + (y_1-y_2)^2,$$

and the points of intersection between ℓ and c_1 (if they exist in $\mathbf{R}^2$) are given as

$$(x_1,y_1) - \frac{ax_1+by_1+c}{a^2+b^2}(a,b) \pm \sqrt{r^2 - (\frac{ax_1+by_1+c}{a^2+b^2})^2}(b,-a). \tag{8.2}$$

The precise expressions given above are not particularly important for our argument, save to note that these expressions are always *algebraic functions* of the input coordinates such as $x_0, x_1, x_2, y_0, y_1, y_2, a, b, c, a', b', c', r_1, r_2$, defined over the reals $\mathbf{R}$, and that the only algebraic operations needed here besides the arithmetic operations of addition, subtraction, multiplication, and division is the square root operation. Thus, we see that any particular construction of, say, a line ℓ from a configuration (A,B,C) will locally be an algebraic function of C (recall that we have already fixed A, B), and this definition can be extended until one reaches a degeneracy (two points, lines, or circles collide, two curves become tangent, or two lines become parallel); however, this degeneracy only occurs in an proper real algebraic set of configurations, and in particular for C in a dimension zero subset of the circle S^1.

These degeneracies are annoying because they disconnect the circle S^1, and can potentially block off large regions of that circle for which the construction is not even defined (because two circles stop intersecting, or a circle and line stop intersecting, in $\mathbf{R}^2$, due to the lack of a real square root for negative numbers). To fix this, we move now from the real plane $\mathbf{R}^2$ to the complex plane $\mathbf{C}^2$. Note that the algebraic definitions of a line and a circle continue to make perfect sense in $\mathbf{C}^2$ (with coefficients such as a, b, c, x_0, y_0, r now allowed to be complex numbers instead of real numbers), and the algebraic intersection formulae given previously continue to make sense in the complex setting. The point C now is allowed to range in the complex circle $S^1_{\mathbf{C}} = \{(x,y) \in \mathbf{C} : x^2 + y^2 = 1\}$, which is a Riemann surface

(conformal to the *Riemann sphere* $\mathbf{C} \cup \infty$ after stereogrpahic projection). Furthermore, because all non-zero complex numbers have square roots, any given construction that was valid for at least one configuration is now valid (though possibly multi-valued) as an algebraic function on $S^1_{\mathbf{C}}$ outside of a dimension zero set of singularities, i.e. outside of a finite number of exceptional values of C. But note now that these singularities do not disconnect the complex circle $S^1_{\mathbf{C}}$, which has topological dimension two instead of one.

As mentioned earlier, a line ℓ given by such a construction may or may not trisect the original angle $\angle BAC$. But this trisection property can be expressed algebraically (e.g. using the triple angle formulae from trigonometry, or by building rotation matrices), and in particular makes sense over $\mathbf{C}$. Thus, for any given construction of a line ℓ, the set of C in $S^1_{\mathbf{C}}$ for which the construction is non-degenerate and trisects $\angle BAC$ is a *constructible set* (a boolean combination of algebraic sets). But $S^1_{\mathbf{C}}$ is an irreducible one-dimensional complex variety. As such, the aforementioned set of C is either *generic* (the complement of a dimension one algebraic set), or has dimension at most one. (Here we are implicitly using the fundamental theorem of algebra, because the basic dimension theory of algebraic geometry only works properly over algebraically closed fields.)

On the other hand, there are at most countably many constructions, and by hypothesis, for each choice of C in $S^1_{\mathbf{C}}$, at least one of these constructions has to trisect the angle. Applying the *Baire category theorem* (or countable additivity of Lebesgue measure, or using the algebraic geometry fact that an algebraic variety over an uncountable field cannot be covered by the union of countably many algebraic sets of smaller dimension), we conclude that there is a single construction which trisects the angle $\angle BAC$ for a generic choice of C, i.e. for all C in $S^1_{\mathbf{C}}$ outside of a finite set of points, there is a construction, which amongst its multiple possible values, is able to output at least one line ℓ that trisects $\angle BAC$.

Now one performs monodromy. Suppose we move C around a closed loop in $S^1_{\mathbf{C}}$ that avoids all points of degeneracy. Then all the other points, lines, and circles constructed from A, B, C can be continuously extended from an initial configuration as discussed earlier, with each such object tracing out its own path in its own configuration space. Because of the presence of square roots in constructions such as the intersection (8.1) between two circles, or the intersection (8.2) between a circle and a line, these constructions may map a closed loop to an open loop; but because the square root function forms a double cover of $\mathbf{C}\backslash 0$, we see that any closed loop in $\mathbf{C}\backslash 0$, if doubled, will continue to be a closed loop upon taking a square root. (Alternatively, one can argue geometrically rather than algebraically, noting that in the intersection of (say) two non-degenerate circles c_1, c_2, there are only two

possible choices for the intersection point of these two circles, and so if one performs monodromy along a loop of possible pairs (c_1, c_2) of circles, either these two choices return to where they initially started, or are swapped; so if one doubles the loop, one must necessarily leave the intersection points unchanged.) Iterating this, we see that any object constructed by straightedge and compass from A, B, C must have period 2^k for some power of two 2^k, in the sense that if one iterates a loop of C in $S^1_{\mathbf{C}}$ avoiding degenerate points 2^k times, the object must return to where it started. (In more algebraic terminology: the monodromy group must be a 2-group.)

Now, one traverses C along a slight perturbation of a single rotation of the real unit circle S^1, taking a slight detour around the finite number of degeneracy points one encounters along the way. Since ℓ has to trisect the angle $\angle ABC$ at each of these points, while varying continuously with C, we see that when C traverses a full rotation, ℓ has only traversed one third of a rotation (or two thirds, depending on which trisection one obtained), and so the period of ℓ must be a multiple of three; but this contradicts Lemma 8.1.2, and the claim follows.

8.2. Elliptic curves and Pappus's theorem

An *algebraic (affine) plane curve* of degree d over some field k is a curve γ of the form

$$\gamma = \{(x, y) \in k^2 : P(x, y) = 0\}$$

where P is some non-constant polynomial of degree d. Examples of low-degree plane curves include

- *Degree* 1 *(linear) curves* $\{(x, y) \in k^2 : ax + by = c\}$, which are simply the *lines*;
- *Degree* 2 *(quadric) curves* $\{(x, y) \in k^2 : ax^2 + bxy + cy^2 + dx + ey + f = 0\}$, which (when $k = \mathbf{R}$) include the classical *conic sections* (i.e. ellipses, hyperbolae, and parabolae), but also include the reducible example of the union of two lines; and
- *Degree* 3 *(cubic) curves* $\{(x, y) \in k^2 : ax^3 + bx^2y + cxy^2 + dy^3 + ex^2 + fxy + gy^2 + hx + iy + j = 0\}$, which include the *elliptic curves* $\{(x, y) \in k^2 : y^2 = x^3 + ax + b\}$ (with non-zero *discriminant* $\Delta := -16(4a^3 + 27b^2)$, so that the curve is smooth) as examples (ignoring some technicalities when k has characteristic two or three), but also include the reducible examples of the union of a line and a conic section, or the union of three lines.
- etc.

Algebraic affine plane curves can also be extended to the projective plane $\mathbb{P}k^2 = \{[x, y, z] : (x, y, z) \in k^3 \backslash 0\}$ by homogenising the polynomial. For

instance, the affine quadric curve $\{(x,y) \in k^2 : ax^2+bxy+cy^2+dx+ey+f = 0\}$ would become $\{[x,y,z] \in \mathbb{P}k^2 : ax^2 + bxy + cy^2 + dxz + eyz + fz^2 = 0\}$.

One of the fundamental theorems about algebraic plane curves is *Bézout's theorem*, which asserts that if a degree d curve γ and a degree d' curve γ' have no common component, then they intersect in at most dd' points (and if the underlying field k is algebraically closed, one works projectively, and one counts intersections with multiplicity, they intersect in *exactly* dd' points). Thus, for instance, two distinct lines intersect in at most one point; a line and a conic section intersect in at most two points; two distinct conic sections intersect in at most four points; a line and an elliptic curve intersect in at most three points; two distinct elliptic curves intersect in at most nine points; and so forth. Bézout's theorem is discussed further in Section 8.4.

From linear algebra we also have the fundamental fact that one can build algebraic curves through various specified points. For instance, for any two points A_1, A_2 one can find a line $\{(x,y) : ax + by = c\}$ passing through the points A_1, A_2, because this imposes two linear constraints on three unknowns a, b, c and is thus guaranteed to have at least one solution. Similarly, given any five points $A_1, \ldots, A_5$, one can find a quadric curve passing through these five points (though note that if three of these points are collinear, then this curve cannot be a conic thanks to Bézout's theorem, and is thus necessarily reducible to the union of two lines); given any nine points $A_1, \ldots, A_9$, one can find a cubic curve going through these nine points; and so forth. This simple observation is one of the foundational building blocks of the *polynomial method* in combinatorial incidence geometry, discussed for instance in [**Ta2009b**, §1.1].

In the degree 1 case, it is always true that two distinct points A, B determine exactly one line $\overleftrightarrow{AB}$. In higher degree, the situation is a bit more complicated. For instance, five collinear points determine more than one quadric curve, as one can simply take the union of the line containing those five points, together with an arbitrary additional line. Similarly, eight points on a conic section plus one additional point determine more than one cubic curve, as one can take that conic section plus an arbitrary line going through the additional point. However, if one places some "general position" hypotheses on these points, then one can recover uniqueness. For instance, given five points, no three of which are collinear, there can be at most one quadric curve that passes through these points (because these five points cannot lie on the union of two lines, and by Bézout's theorem they cannot simultaneously lie on two distinct conic sections).

For cubic curves, the situation is more complicated still. Consider for instance two distinct cubic curves $\gamma_0 = \{P_0(x,y) = 0\}$ and $\gamma_\infty = \{P_\infty(x,y) = 0\}$ that intersect in precisely nine points $A_1, \ldots, A_9$ (note from Bézout's

theorem that this is an entirely typical situation). Then there is in fact an entire one-parameter family of cubic curves that pass through these points, namely the curves $\gamma_t = \{P_0(x,y) + tP_\infty(x,y) = 0\}$ for any $t \in k \cup \{\infty\}$ (with the convention that the constraint $P_0 + tP_\infty = 0$ is interpreted as $P_\infty = 0$ when $t = \infty$).

In fact, these are the only cubics that pass through these nine points, or even through eight of the nine points. More precisely, we have the following useful fact, known as the *Cayley-Bacharach theorem* (although the version given here is actually due to Chasles [**Ch1885**]):

Proposition 8.2.1 (Cayley-Bacharach theorem)**.** *Let $\gamma_0 = \{P_0(x,y) = 0\}$ and $\gamma_\infty = \{P_\infty(x,y) = 0\}$ be two cubic curves that intersect (over some algebraically closed field k) in precisely nine distinct points $A_1, \ldots, A_9 \in k^2$. Let P be a cubic polynomial that vanishes on eight of these points (say $A_1, \ldots, A_8$). Then P is a linear combination of P_0, P_∞, and in particular vanishes on the ninth point A_9.*

Proof. We use an argument from [**Hu2004**]. We assume for contradiction that there is a cubic polynomial P that vanishes on $A_1, \ldots, A_8$, but is not a linear combination of P_0 and P_∞.

We first make some observations on the points $A_1, \ldots, A_9$. No four of these points can be collinear, because then by Bézout's theorem, P_0 and P_∞ would both have to vanish on this line, contradicting the fact that γ_0, γ_∞ meet in at most nine points. For similar reasons, no seven of these points can lie on a quadric curve.

One consequence of this is that any five of the $A_1, \ldots, A_9$ determine a unique quadric curve σ. The existence of the curve follows from linear algebra as discussed previously. If five of the points lie on two different quadric curves σ, σ', then by Bezout's theorem, they must share a common line; but this line can contain at most three of the five points, and the other two points determine uniquely the other line that is the component of both σ and σ', and the claim follows.

Now suppose that three of the first eight points, say A_1, A_2, A_3, are collinear, lying on a line ℓ. The remaining five points $A_4, \ldots, A_8$ do not lie on ℓ, and determine a unique quadric curve σ by the previous discussion. Let B be another point on ℓ, and let C be a point that does not lie on either ℓ or σ. By linear algebra, one can find a non-trivial linear combination $Q = aP + bP_0 + cP_\infty$ of P, P_0, P_∞ that vanishes at both B and C. Then Q is a cubic polynomial that vanishes on the four collinear points A_1, A_2, A_3, B and thus vanishes on ℓ, thus the cubic curve defined by Q consists of ℓ and a quadric curve. This curve passes through $A_4, \ldots, A_8$ and thus equals σ.

But then C does not lie on either ℓ or σ despite being a vanishing point of Q, a contradiction. Thus, no three points from $A_1, \ldots, A_8$ are collinear.

In a similar vein, suppose next that six of the first eight points, say $A_1, \ldots, A_6$, lie on a quadric curve σ; as no three points are collinear, this quadric curve cannot be the union of two lines, and is thus a conic section. The remaining two points A_7, A_8 determine a unique line $\ell = \overleftrightarrow{A_7 A_8}$. Let B be another point on σ, and let C be another point that does not lie on either ℓ and σ. As before, we can find a non-trivial cubic $Q = aP + bP_0 + cP_\infty$ that vanishes at both B, C. As Q vanishes at seven points of a conic section σ, it must vanish on all of σ, and so the cubic curve defined by Q is the union of σ and a line that passes through A_7 and A_8, which must necessarily be ℓ. But then this curve does not pass through C, a contradiction. Thus no six points in $A_1, \ldots, A_8$ lie on a quadric curve.

Finally, let ℓ be the line through the two points A_1, A_2, and σ the quadric curve through the five points $A_3, \ldots, A_7$; as before, σ must be a conic section, and by the preceding paragraphs we see that A_8 does not lie on either σ or ℓ. We pick two more points B, C lying on ℓ but not on σ. As before, we can find a non-trivial cubic $Q = aP + bP_0 + cP_\infty$ that vanishes on B, C; it vanishes on four points on ℓ and thus Q defines a cubic curve that consists of ℓ and a quadric curve. The quadric curve passes through $A_3, \ldots, A_7$ and is thus σ; but then the curve does not pass through A_8, a contradiction. This contradiction finishes the proof of the proposition. □

I recently learned of this proposition and its role in unifying many incidence geometry facts concerning lines, quadric curves, and cubic curves. For instance, we can recover the proof of the classical theorem of Pappus:

Theorem 8.2.2 (Pappus' theorem)**.** *Let ℓ, ℓ' be two distinct lines, let A_1, A_2, A_3 be distinct points on ℓ that do not lie on ℓ', and let B_1, B_2, B_3 be distinct points on ℓ' that do not lie on ℓ. Suppose that for $ij = 12, 23, 31$, the lines $\overleftrightarrow{A_i B_j}$ and $\overleftrightarrow{A_j B_i}$ meet at a point C_{ij}. Then the points C_{12}, C_{23}, C_{31} are collinear.*

Proof. We may assume that C_{12}, C_{23} are distinct, since the claim is trivial otherwise.

Let γ_0 be the union of the three lines $\overleftrightarrow{A_1 B_2}$, $\overleftrightarrow{A_2 B_3}$, and $\overleftrightarrow{A_3 B_1}$ (the purple lines in the first figure), let γ_1 be the union of the three lines $\overleftrightarrow{A_2 B_1}$, $\overleftrightarrow{A_3 B_2}$, and $\overleftrightarrow{A_1 B_3}$ (the dark blue lines), and let γ be the union of the three lines ℓ, ℓ', and $\overleftrightarrow{C_{12} C_{23}}$ (the other three lines). By construction, γ_0 and γ_1 are cubic curves with no common component that meet at the nine points $A_1, A_2, A_3, B_1, B_2, B_3, C_{12}, C_{23}, C_{31}$. Also, γ is a cubic curve that passes through the first eight of these points, and thus also passes through the

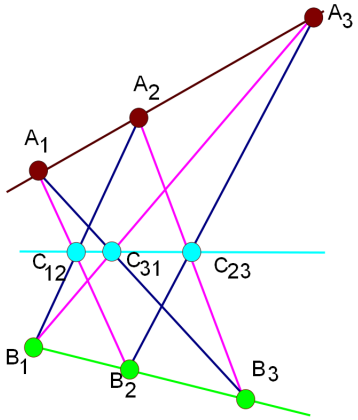

Figure 11. Pappus' theorem.

ninth point C_{31}, by the Cayley-Bacharach theorem. The claim follows (note that C_{31} cannot lie on ℓ or ℓ'). □

The same argument gives the closely related theorem of Pascal:

Theorem 8.2.3 (Pascal's theorem)**.** *Let $A_1, A_2, A_3, B_1, B_2, B_3$ be distinct points on a conic section σ. Suppose that for $ij = 12, 23, 31$, the lines $\overleftrightarrow{A_iB_j}$ and $\overleftrightarrow{A_jB_i}$ meet at a point C_{ij}. Then the points C_{12}, C_{23}, C_{31} are collinear.*

Proof. Repeat the proof of Pappus' theorem, with σ taking the place of $\ell \cup \ell'$. (Note that as any line meets σ in at most two points, the C_{ij} cannot lie on σ.) □

One can view Pappus's theorem as the degenerate case of Pascal's theorem, when the conic section degenerates to the union of two lines.

Finally, Proposition 8.2.1 gives the associativity of the elliptic curve group law:

Theorem 8.2.4 (Associativity of the elliptic curve law)**.** *Let $\gamma := \{(x, y) \in k^2 : y^2 = x^3+ax+b\} \cup \{O\}$ be a (projective) elliptic curve, where $O := [0, 1, 0]$ is the point at infinity on the y-axis, and the discriminant $\Delta := -16(4a^3 + 27b^2)$ is non-zero. Define an addition law $+$ on γ by defining $A+B$ to equal $-C$, where C is the unique point on γ collinear with A and B (if A, B are disjoint) or tangent to A (if $A = B$), and $-C$ is the reflection of C through*

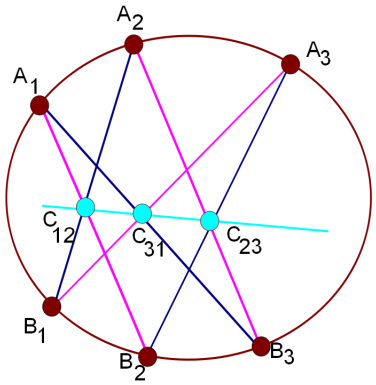

Figure 12. Pascal's theorem.

the x-axis (thus $C, -C, O$ are collinear), with the convention $-O = O$. Then $+$ gives γ the structure of an abelian group with identity O and inverse $-$.

Proof. It is clear that O is the identity for $+$, $-$ is an inverse, and $+$ is abelian. The only non-trivial assertion is associativity: $(A + B) + C = A + (B + C)$. By a perturbation (or Zariski closure) argument, we may assume that we are in the generic case when $O, A, B, C, A+B, B+C, -(A+B), -(B+C)$ are all distinct from each other and from $-((A+B)+C), -(A+(B+C))$. (Here we are implicitly using the smoothness of the elliptic curve, which is guaranteed by the hypothesis that the discriminant is non-zero.)

Let γ' be the union of the three lines $\overleftrightarrow{AB}$, $\overleftrightarrow{C(A+B)}$, and $\overleftrightarrow{O(B+C)}$ (the purple lines), and let γ'' be the union of the three lines $\overleftrightarrow{O(A+B)}$, $\overleftrightarrow{BC}$, and $\overleftrightarrow{A(B+C)}$ (the green lines). Observe that γ' and γ are cubic curves with no common component that meet at the nine distinct points $O, A, B, C, A+B, -(A+B), B+C, -(B+C), -((A+B)+C)$. The cubic curve γ'' goes through the first eight of these points, and thus (by Proposition 8.2.1) also goes through the ninth point $-((A+B)+C)$. This implies that the line through A and $B+C$ meets γ in both $-(A+(B+C))$ and $-((A+B)+C)$, and so these two points must be equal, and so $(A + B) + C = A + (B + C)$ as required. □

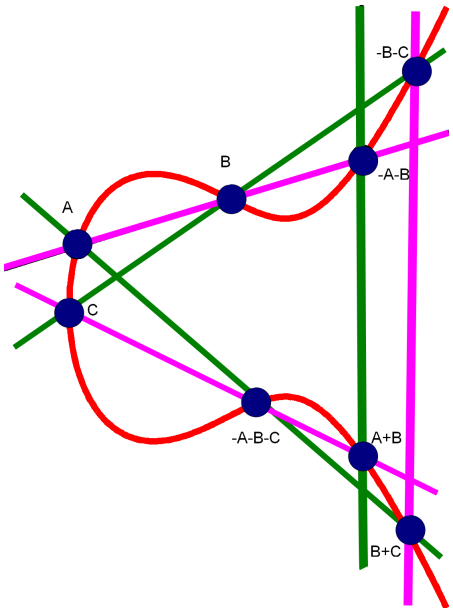

Figure 13. Associativity of the elliptic curve group law.

One can view Pappus's theorem and Pascal's theorem as a degeneration of the associativity of the elliptic curve law, when the elliptic curve degenerates to three lines (in the case of Pappus) or the union of one line and one conic section (in the case of Pascal's theorem).

8.3. Lines in the Euclidean group SE(2)

In the recent breakthrough of Guth and Katz [**GuKa2010**] (discussed at [**Ta2011d**, §3.9]) on the Erdős distance problem, one of the main tools used in the proof (building upon the earlier work of Elekes and Sharir [**ElSh2011**]) was the observation (dating back to [**St1891**]) that the incidence geometry of the Euclidean group SE(2) of rigid motions of the plane was almost identical to that of lines in the Euclidean space $\mathbf{R}^3$:

Proposition 8.3.1. *One can identify a (Zariski-)dense portion of* SE(2) *with* $\mathbf{R}^3$*, in such a way that for any two points* A, B *in the plane* $\mathbf{R}^2$*, the set* $l_{AB} := \{R \in \mathrm{SE}(2) : RA = B\}$ *of rigid motions mapping* A *to* B *forms a line in* $\mathbf{R}^3$*.*

Proof. A rigid motion is either a translation or a rotation, with the latter forming a Zariski-dense subset of SE(2). Identify a rotation R in SE(2) by an angle θ with $|\theta| < \pi$ around a point P with the element $(P, \cot\frac{\theta}{2})$ in $\mathbf{R}^3$. (Note that such rotations also form a Zariski-dense subset of SE(2).)

Elementary trigonometry then reveals that if R maps A to B, then P lies on the perpendicular bisector of AB, and depends in a linear fashion on $\cot \frac{\theta}{2}$ (for fixed A, B). The claim follows. □

As seen from the proof, this proposition is an easy (though *ad hoc*) application of elementary trigonometry, but it was still puzzling to me why such a simple parameterisation of the incidence structure of SE(2) was possible. Certainly it was clear from general algebraic geometry considerations that *some* bounded-degree algebraic description was available, but why would the l_{AB} be expressible as lines and not as, say, quadratic or cubic curves?

In this section I would like to record some observations arising from discussions with Jordan Ellenberg, Jozsef Solymosi, and Josh Zahl which give a more conceptual (but less elementary) derivation of the above proposition that avoids the use of *ad hoc* coordinate transformations such as $R \mapsto (P, \cot \frac{\theta}{2})$. The starting point is to view the Euclidean plane $\mathbf{R}^2$ as the scaling limit of the sphere S^2 (a fact which is familiar to all of us through the geometry of the Earth), which makes the Euclidean group SE(2) a scaling limit of the rotation group SO(3). The latter can then be lifted to a double cover, namely the *spin group* Spin(3). This group has a natural interpretation as the unit *quaternions*, which is isometric to the unit sphere S^3. The analogue of the lines l_{AB} in this setting become *great circles* on this sphere; applying a *projective transformation*, one can map S^3 to $\mathbf{R}^3$ (or more precisely to the *projective space* $\mathbb{P}^3$), at whichi point the great circles become lines. This gives a proof of Proposition 8.3.1.

Details of the correspondence are provided below the fold. One by-product of this analysis, incidentally, is the observation that the Guth-Katz bound $g(N) \gg N/\log N$ for the Erdos distance problem in the plane $\mathbf{R}^2$, immediately extends with almost no modification to the sphere S^2 as well (i.e. any N points in S^2 determine $\gg N/\log N$ distances), as well as to the hyperbolic plane H^2. (The latter observation has since been applied in [**IoRoRu2011**].)

8.3.1. Euclidean geometry as the scaling limit of spherical geometry. *Euclidean geometry* and *spherical geometry* are examples of *Kleinian geometries*: the geometry of a space X with a group of symmetries G acting transitively on it. In the case of Euclidean plane geometry, the space X is the plane $\mathbf{R}^2$ and the symmetry group is the special Euclidean group $\mathrm{SE}(2) = \mathrm{SO}(2) \ltimes \mathbf{R}^2$; in the case of spherical geometry, the space X is the unit sphere S^2 and the symmetry group is the special orthogonal group SO(3). According to the Kleinian way of thinking (as formalised by the *Erlangen program*), explicit coordinates on X should be avoided if possible,

with a preference instead for only using concepts (e.g. congruence, distance, angle) that are invariant with respect to the group of symmetries G.

As we all know from the geometry of the Earth (and the Greek root *geometria* literally means "Earth measurement"), the geometry of the sphere S^2 resembles the geometry of the plane $\mathbf{R}^2$ at scales that are small compared to the radius of the sphere. There are at least two ways to make this intuitive fact more precise. One is to make the radius R of the sphere go to infinity, and perform a suitable limit (e.g. a *Gromov-Hausdorff limit*). A dual approach is to keep the radius of the sphere fixed (e.g. considering only the unit sphere), but making the scale ε being considered on the sphere shrink to zero. The two approaches are of course equivalent, but we will consider the latter.

Thus, we view S^2 as the unit sphere in $\mathbf{R}^3$. With an eye to using the quaternionic number system later on, we will denote the standard basis of $\mathbf{R}^3$ as i, j, k, thus in particular i is a point on the sphere S^2 which we will view as an "origin" for this sphere. The tangent plane to S^2 at this point is then

$$\{i + yj + zk : y, z \in \mathbf{R}^2\}.$$

This plane is tangent to the sphere to second order. In particular, if $(y, z) \in \mathbf{R}^2$, and $\varepsilon > 0$ is a small parameter (which we think of as going to zero eventually), then we can find a point on S^2 of the form $i + \varepsilon yj + \varepsilon zk + O(\varepsilon^2)$. (If one wishes, one can enforce the $O(\varepsilon^2)$ error to lie in the i direction, in order to make the identification uniquely well-defined, although this is not strictly necessary for the discussion below.) Thus, we can view the ε-neighbourhood of the origin i as being approximately identifiable with a bounded neighbourhood of the origin 0 in the plane $\mathbf{R}^2$ via the identification

$$(y, z) \mapsto i + \varepsilon yj + \varepsilon zk + O(\varepsilon^2).$$

With this identification, one can see various structures in spherical geometry correspond (up to errors of $O(\varepsilon)$) to analogous structures in planar geometry. For instance, a great circle in S^2 is of the form

$$\{p \in S^2 : p \cdot \omega = 0\}$$

for some $\omega \in S^2$, where $\cdot$ is the usual dot product. In order for this great circle to intersect the $O(\varepsilon)$ neighbourhood of the origin i, one must have $i \cdot \omega = O(\varepsilon)$, and so we have

$$\omega = \varepsilon ai + (\cos\theta)j + (\sin\theta)k + O(\varepsilon^2)$$

for some bounded quantity a and some angle θ. If one then restricts the great circle to points $p = i + \varepsilon yj + \varepsilon zk + O(\varepsilon^2)$, the constraint $p \cdot \omega = 0$ then becomes

$$a + (\cos\theta)y + (\sin\theta)z = O(\varepsilon),$$

which is within $O(\varepsilon)$ of the equation of a typical line in $\mathbf{R}^2$,

$$a + (\cos\theta)y + (\sin\theta)z = 0.$$

This formalises the geometrically intuitive fact that great circles resemble lines at small scales.

Remark 8.3.2. One can also adopt a more "intrinsic" Riemannian geometry viewpoint to see $\mathbf{R}^2$ as the limit of rescaled versions of S^2. Indeed, for each real number κ, there is a unique (up to isometry) simply connected Riemannian surface S_κ of constant *scalar curvature* κ. For $\kappa > 0$, this is the unit sphere S^2 (rescaled by $\sqrt{\kappa/2}$); for $\kappa = 0$, this is the Euclidean plane $\mathbf{R}^2$; and for $\kappa < 0$, it is the *hyperbolic plane* H^2 (rescaled by $\sqrt{|\kappa|/2}$). Sending κ to zero, we thus see the sphere (or hyperbolic plane) converging to the Euclidean plane.

We now apply a similar analysis to a rotation matrix $R \in \mathrm{SO}(3)$ acting on the unit sphere S^2. In order for this rotation matrix to map an $O(\varepsilon)$ neighbourhood of the origin i to another such neighbourhood, the rotation matrix R must be of the form $R = (1 + \varepsilon T + O(\varepsilon^2))S$, where S is a rotation that fixes i, thus

$$S(xi + yj + zk) = xi + (y\cos\theta - z\sin\theta)j + (y\sin\theta + z\cos\theta)k$$

for some angle θ, and T is an infinitesimal rotation (i.e. an element of the Lie algebra $\mathfrak{so}(3)$), thus $Ti = aj + bk$ for some reals a, b. We then have

$$\begin{aligned} R(i + \varepsilon yj + \varepsilon zk + O(\varepsilon^2)) &= i + \varepsilon(y\cos\theta - z\sin\theta + a)j \\ &\quad + \varepsilon(y\sin\theta + z\cos\theta + b) + O(\varepsilon^2), \end{aligned}$$

so in (y, z) coordinates, R becomes the map

$$(y, z) \mapsto (y\cos\theta - z\sin\theta + a, y\sin\theta + z\cos\theta + b) + O(\varepsilon),$$

which is within ε of the Euclidean rigid motion

$$(y, z) \mapsto (y\cos\theta - z\sin\theta + a, y\sin\theta + z\cos\theta + b).$$

Thus we see the Euclidean rigid motion group SE(2) emerging as a scaling limit of the orthogonal rotation group SO(3) (or alternatively, as the normal bundle of the stabiliser of the origin, which is a copy of SO(2)).

Remark 8.3.3. One can also analyse the situation from a Lie algebra perspective. As is well known, one can equip the three-dimensional Lie algebra $\mathfrak{so}(3)$ with a basis X, Y, Z obeying the commutation relations

$$[X, Y] = Z; [Y, Z] = X; [Z, X] = Y,$$

corresponding to infinitesimal rotations around the x, y, z axes respectively. If we then rescale $X_\varepsilon := X, Y_\varepsilon := \varepsilon Y, Z_\varepsilon := \varepsilon Z$ (which morally corresponds

to looking at rotation matrices that almost fix i, as above), the commutation relations rescale to

$$[X_\varepsilon, Y_\varepsilon] = Z_\varepsilon; [Y_\varepsilon, Z_\varepsilon] = \varepsilon^2 X_\varepsilon; [Z_\varepsilon, X_\varepsilon] = Y_\varepsilon.$$

Sending $\varepsilon \to 0$, the Lie algebra degenerates to the solvable Lie algebra

$$[X_0, Y_0] = Z_0; [Y_0, Z_0] = 0; [Z_0, X_0] = Y_0,$$

which is the Lie algebra $\mathfrak{se}(2)$ of the Euclidean group SE(2).

There is an exact analogue of this phenomenon for the isometry group $\mathrm{SO}(2,1) \equiv SL_2(\mathbf{R})$ of the hyperbolic plane H^2 (which one can think of as one sheet of the unit sphere in Minkowski space $\mathbf{R}^{2+1}$, just as S^2 is the unit sphere in $\mathbf{R}^3$). The Lie algebra here can be equipped with a basis X, Y, Z (which one can interpret as infinitesimal rotations and Lorentz boosts in Minkowski space) with the relations

$$[X, Y] = Z; [Y, Z] = -X; [Z, X] = Y,$$

and the same scaling argument as before gives SE(2) as a scaling limit of SO(2, 1).

8.3.2. Lifting to the quaternions. The *quaternions* are a number system, defined formally as the set $\mathbf{H}$ of numbers of the form

$$t + xi + yj + zk$$

with $t, x, y, z \in \mathbf{R}$. This a four-dimensional vector space; it can be turned into an algebra (and into a *division ring*) by enforcing Hamilton's relations

$$i^2 = j^2 = k^2 = ijk = -1.$$

The quaternions also come equipped with a conjugation operation

$$(t + xi + yj + zk)^* := t - xi - yj - zk$$

and a norm

$$|\alpha| := (\alpha\alpha^*)^{1/2} = (\alpha^*\alpha)^{1/2},$$

thus

$$|(t + xi + yj + zk)| = \sqrt{t^2 + x^2 + y^2 + z^2}.$$

The conjugation operation is an *anti-automorphism*, and the norm is multiplicative: $|\alpha\beta| = |\alpha||\beta|$. The quaternions also have a trace

$$\mathrm{tr}(t + xi + yj + zk) = t$$

(in particular, $\mathrm{tr}(\alpha^*) = \mathrm{tr}(\alpha)$ and $\mathrm{tr}(\alpha\beta) = \mathrm{tr}(\beta\alpha)$), giving rise to a dot product

$$\alpha \cdot \beta := \mathrm{tr}(\alpha\beta^*)$$

which (together with the norm) gives a Hilbert space structure on the quaternions.

The unit sphere $S^3 = \{\alpha \in \mathbf{H} : \alpha\alpha^* = 1\}$ of the quaternions forms a group, which is a model for the *spin group* Spin(3) (thus giving rise to an interpretation of the quaternions $\mathbf{H}$, which are of course acted upon by S^3 by left-multiplication, as *spinors* for this group). This group acts on itself isometrically by conjugation, with an element $\alpha \in S^3 \equiv \mathrm{Spin}(3)$ mapping $\beta \in S^3$ to $\alpha\beta\alpha^*$. As $\alpha\alpha^* = 1$, this action preserves the quaternionic identity 1, and thus preserves the orthogonal complement $\{xi + yj + zk : x^2+y^2+z^2\}$ of that identity in S^3, which we can of course identify with S^2. Thus $S^3 \equiv \mathrm{Spin}(3)$ acts on S^2 isometrically by conjugation, thus providing a map from Spin(3) to SO(3). One can verify that this map is surjective (indeed, conjugation by the quaternion $e^{i\theta}$ corresponds to a rotation around the i-axis by 2θ, and similarly for rotations around other axes) and is a double cover (since the center of S^3 is $\{-1, +1\}$), with the pre-image of any rotation in SO(3) being a pair $\{\alpha, -\alpha\}$ of diametrically opposed points in S^3. Thus we see that SO(3) can be identified (in either the topological or Riemannian geometrical senses) with the projective sphere $S^3/\pm \equiv \mathbf{P}^3$. As SO(3) acts transitively on S^2, we see that $S^3 \equiv \mathrm{Spin}(3)$ does also.

The stabiliser $l_{AA} := \{\alpha \in S^3 : \alpha A\alpha^* = A\}$ of a point $A \in S^2$ is easily seen to be a great circle in S^3 (being the intersection of S^3 with the center of α, which is a plane). For instance, the stabiliser l_{ii} of the origin i is the circle $\{e^{i\theta} : \theta \in \mathbf{R}\}$. (This, incidentally, gives an explicit geometric manifestation of the *Hopf fibration*.) By transitivity (and the isometric nature of the S^3 action), we conclude that the sets $l_{AB} := \{\alpha \in S^3 : \alpha A\alpha^* = B\}$ are also great circles in S^3 for any pair of points A, B. Conversely, as all great circles are isometric to each other, we see that all great circles are of the form l_{AB}. One also sees that two great circles l_{AB}, l_{CD} coincide only when A, C have the same stabiliser, and when B, D have the same stabiliser, which forces C, D to either equal A, B, or $-A, -B$.

Remark 8.3.4. Using a projective transformation, one can identify (a hemisphere of) S^3 with $\mathbf{R}^3$, with (most) great circles becoming lines in $\mathbf{R}^3$. Thus, we see that the incidence geometry of the l_{AB} in S^3 is essentially equivalent to the incidence geometry of lines in $\mathbf{R}^3$. Because of this, the Guth-Katz argument to establish the bound $g(N) \gg N/\log N$ for the number of distances determined by N points in the plane $\mathbf{R}^2$, also extends to N points in the sphere S^2. Indeed, as in the Guth-Katz paper, it suffices to show that the number of congruent line segments AB, CD in these N points is $O(N^3 \log N)$. For each such pair of line segments, there is a unique element of SO(3) (and thus two antipodal elements of $S^3 \equiv \mathrm{Spin}(3)$) that maps AB to CD; these two antipodal points are also the intersection of l_{AC} with l_{BD}. Applying the projective transformation, one arrives at exactly the same incidence problem between points and lines considered by Guth-Katz (and in

particular, one can apply [**GuKa2010**, Theorems 2.4, 2.5] as a black box, after verifying that at most $O(N)$ lines of the form l_{AB} project into a plane or regulus, which is proven in the S^2 case in much the same way as it is in the $\mathbf{R}^2$ case). We omit the details.

A similar argument (changing the signatures in various metrics, and in the Clifford algebra underlying the quaternions) also allows one to establish the same results in the hyperbolic plane H^2; again, we omit the details.

If we restrict attention to an ε-neighbourhood of the origin i in the sphere S^2, and similarly restrict to an ε-neighbourhood of the stabiliser of i in the spin group $S^3 \equiv \mathrm{Spin}(3)$, we can use the correspondences from the previous section to convert S^2 into $\mathbf{R}^2$ in the limit, and $\mathrm{Spin}(3)$ in the limit into a double cover of the rotation group $\mathrm{SE}(2)$ (which ends up just being isomorphic to $\mathrm{SE}(2)$ again). The great circles l_{AB} in $\mathrm{Spin}(3)$ then, in the limit, become the analogous sets $l_{AB} = \{R \in \mathrm{SE}(2) : RA = B\}$ in $\mathrm{SE}(2)$, and the above correspondences can then be used to map (most of) $\mathrm{SE}(2)$ to $\mathbf{R}^3$, and (most) l_{AB} to lines, giving Proposition 8.3.1.

Remark 8.3.5. The results in this section can also be interpreted using the language of Clifford algebra; see [**Gu2011**].

8.4. Bezout's inequality

Classically, the fundamental object of study in *algebraic geometry* is the solution set

$$V = V_{P_1,\dots,P_m} := \{(x_1,\dots,x_d) \in k^d : P_1(x_1,\dots,x_d) = \dots = P_m(x_1,\dots,x_d) = 0\} \tag{8.3}$$

to multiple algebraic equations

$$P_1(x_1,\dots,x_d) = \dots = P_m(x_1,\dots,x_d) = 0$$

in multiple unknowns $x_1,\dots,x_d$ in a field k, where the $P_1,\dots,P_m : k^d \to k$ are polynomials of various degrees $D_1,\dots,D_m$. We adopt the classical perspective of viewing V as a set (and specifically, as an *algebraic set*), rather than as a scheme. Without loss of generality we may order the degrees in non-increasing order:

$$D_1 \geq D_2 \geq \dots \geq D_m \geq 1.$$

We can distinguish between the *underdetermined case* $m < d$, when there are more unknowns than equations; the *determined case* $m = d$ when there are exactly as many unknowns as equations; and the *overdetermined case* $m > d$, when there are more equations than unknowns.

Experience has shown that the theory of such equations is significantly simpler if one assumes that the underlying field k is *algebraically closed*,

and so we shall make this assumption throughout the rest of this section. In particular, this covers the important case when $k = \mathbf{C}$ is the field of complex numbers (but it does *not* cover the case $k = \mathbf{R}$ of real numbers - see below).

From the general "soft" theory of algebraic geometry, we know that the algebraic set V is a union of finitely many algebraic varieties, each of dimension at least $d - m$, with none of these components contained in any other. In particular, in the underdetermined case $m < d$, there are no zero-dimensional components of V, and thus V is either empty or infinite.

Now we turn to the determined case $d = m$, where we expect the solution set V to be zero-dimensional and thus finite. Here, the basic control on the solution set is given by *Bezout's theorem*. In our notation, this theorem states the following:

Theorem 8.4.1 (Bezout's theorem)**.** *Let $d = m = 2$. If V is finite, then it has cardinality at most $D_1 D_2$.*

This result can be found in any introductory algebraic geometry textbook; it can for instance be proven using the classical tool of *resultants*. The solution set V will be finite when the two polynomials P_1, P_2 are coprime, but can (and will) be infinite if P_1, P_2 share a non-trivial common factor.

By defining the right notion of multiplicity on V (and adopting a suitably "*scheme*-theoretic" viewpoint), and working in *projective space* rather than affine space, one can make the inequality $|V| \leq D_1 D_2$ an equality. However, for many applications (and in particular, for the applications to combinatorial incidence geometry), the upper bound usually suffices.

Bezout's theorem can be generalised in a number of ways. For instance, the restriction on the finiteness of the solution set V can be dropped by restricting attention to V^0, the union of the zero-dimensional irreducible components of V:

Corollary 8.4.2 (Bezout's theorem, again)**.** *Let $d = m = 2$. Then V^0 has cardinality at most $D_1 D_2$.*

Proof. We factor P_1, P_2 into irreducible factors (using unique factorisation of polynomials). By removing repeated factors, we may assume P_1, P_2 are square-free. We then write $P_1 = Q_1 R$, $P_2 = Q_2 R$ where R is the greatest common divisor of P_1, P_2 and Q_1, Q_2 are coprime. Observe that the zero-dimensional component of $\{P_1 = P_2 = 0\}$ is contained in $\{Q_1 = Q_2 = 0\}$, which is finite from the coprimality of Q_1, Q_2. The claim follows. □

It is also not difficult to use Bezout's theorem to handle the overdetermined case $m > d = 2$ in the plane:

Corollary 8.4.3 (Bezout's theorem, yet again)**.** *Let $m \geq d = 2$. Then V^0 has cardinality at most $D_1 D_2$.*

Proof. We may assume all the P_i are square-free. We write $P_1 = Q_2 R_2$, where Q_2 is coprime to P_2 and R_2 divides P_2 (and also P_1). We then write $R_2 = Q_3 R_3$, where Q_3 is coprime to P_3 and R_3 divides P_3 (and also P_1, P_2). Continuing in this fashion we obtain a factorisation $P_1 = Q_2 Q_3 \ldots Q_m R_m$. One then observes that V^0 is contained in the set $\bigcup_{i=2}^m \{Q_i = P_i = 0\}$, which by Theorem 8.4.1 has cardinality at most $\sum_{i=2}^m \deg(Q_i) D_i$. Since $D_i \leq D_2$ and $\sum_{i=2}^m \deg(Q_i) \leq \deg(P_1) = D_1$, the claim follows. □

Remark 8.4.4. Of course, in the overdetermined case $m > d$ one generically expects the solution set to be empty, but if there is enough degeneracy or numerical coincidence then non-zero solution sets can occur. In particular, by considering the case when $P_2 = \ldots = P_m$ and $D_2 = \ldots = D_m$ we see that the bound $D_1 D_2$ can be sharp in some cases. However, one can do a little better in this situation; by decomposing P_m into irreducible components, for instance, one can improve the upper bound of $D_1 D_2$ slightly to $D_1 D_m$. However, this improvement seems harder to obtain in higher dimensions (see below).

Bezout's theorem also extends to higher dimensions. Indeed, we have

Theorem 8.4.5 (Higher-dimensional Bezout's theorem)**.** *Let $d = m \geq 0$. If V is finite, then it has cardinality at most $D_1 \ldots D_d$.*

This is a standard fact, and can for instance be proved from the more general and powerful machinery of *intersection theory*. A typical application of this theorem is to show that, given a degree D polynomial $P : \mathbf{R}^d \to \mathbf{R}$ over the reals, the number of connected components of $\{x \in \mathbf{R}^d : P(x) \neq 0\}$ is $O(D^d)$. The main idea of the proof is to locate a critical point $\nabla P(x) = 0$ inside each connected component, and use Bezout's theorem to count the number of zeroes of the polynomial map $\nabla P : \mathbf{R}^d \to \mathbf{R}^d$. (This doesn't quite work directly because some components may be unbounded, and because the fibre of ∇P at the origin may contain positive-dimensional components, but one can use truncation and generic perturbation to deal with these issues; see [**SoTa2011**] for further discussion.)

Bezout's theorem can be extended to the overdetermined case as before:

Theorem 8.4.6 (Bezout's inequality)**.** *Let $m \geq d \geq 0$. Then V^0 has cardinality at most $D_1 \ldots D_d$.*

Remark 8.4.7. Theorem 8.4.6 ostensibly only controls the zero-dimensional components of V, but by throwing in a few generic affine-linear forms to the set of polynomials $P_1, \ldots, P_m$ (thus intersecting V with a bunch of generic

hyperplanes) we can also control the total degree of all the i-dimensional components of V for any fixed i. (Again, by using intersection theory one can get a slightly more precise bound than this, but the proof of that bound is more complicated than the arguments given here.)

This time, though, it is a slightly non-trivial matter to deduce Theorem 8.4.6 from Theorem 8.4.5, due to the standard difficulty that the intersection of irreducible varieties need not be irreducible (which can be viewed in some ways as the source of many other related difficulties, such as the fact that not every algebraic variety is a *complete intersection*), and so one cannot evade all irreducibility issues merely by assuming that the original polynomials P_i are irreducible. Theorem 8.4.6 first appeared explicitly in the work of Heintz [**He1983**].

As before, the most systematic way to establish Theorem 8.4.6 is via intersection theory. In this section, though, I would like to give a slightly more elementary argument (essentially due to Schmid [**Sc1995**]), based on generically perturbing the polynomials $P_1, \ldots, P_m$ in the problem ; this method is less powerful than the intersection-theoretic methods, which can be used for a wider range of algebraic geometry problems, but suffices for the narrow objective of proving Theorem 8.4.6. The argument requires some of the "soft" or "qualitative" theory of algebraic geometry (in particular, one needs to understand the semicontinuity properties of preimages of dominant maps), as well as basic linear algebra. As such, the proof is not completely elementary, but it uses only a small amount of algebraic machinery, and as such I found it easier to understand than the intersection theory arguments.

Theorem 8.4.6 is a statement about *arbitrary* polynomials $P_1, \ldots, P_m$. However, it turns out (in the determined case $m = d$, at least) that upper bounds on $|V^0|$ are *Zariski closed* properties, and so it will suffice to establish this claim for *generic* polynomials $P_1, \ldots, P_m$. On the other hand, it is possible to use duality to deduce such upper bounds on $|V^0|$ from a *Zariski open* condition, namely that a certain collection of polynomials are linearly independent. As such, to verify the generic case of this open condition, it suffices to establish this condition for a *single* family of polynomials, such as a family of monomials, in which case the condition can be verified by direct inspection. Thus we see an example of the somewhat strange strategy of establishing the general case from a specific one, using the generic case as an intermediate step.

Remark 8.4.8. There is an important caveat to note here, which is that these theorems only hold for algebraically closed fields, and in particular can

fail over the reals $\mathbf{R}$. For instance, in $\mathbf{R}^3$, the polynomials

$$P(x,y,z) = z$$
$$Q(x,y,z) = z$$
$$R(x,y,z) = (\prod_{j=1}^{N}(x-j))^2 + (\prod_{j=1}^{N}(y-j))^2$$

have degrees $1, 1, 2N$ respectively, but their common zero locus $\{(x,y,0) : x,y \in \{1,\ldots,N\}\}$ has cardinality N^2. In some cases one can safely obtain incidence bounds in $\mathbf{R}$ by embedding $\mathbf{R}$ inside $\mathbf{C}$ (for instance, things are OK if one knows that the zero locus has *complex* dimension zero, and not merely *real* dimension zero), but as the above example shows, one needs to be careful when doing so.

8.4.1. The determined case. We begin by establishing Theorem 8.4.5. Fix $m = d \geq 0$. If one wishes, one can dispose of the trivial case $m = d = 0$ and assume $m = d \geq 1$, although this is not strictly necessary for the argument below.

As mentioned in the introduction, the first idea is to generically perturb the $P_1, \ldots, P_d$. To this end, let W be the vector space $W := W_1 \times \ldots \times W_d$, where W_i is the vector space of all polynomials $P_i : k^d \to k$ of degree at most D_i; thus W is the configuration space of all possible $P_1, \ldots, P_d$. This is a finite-dimensional vector space over k with an explicit dimension $\dim(W)$ which we will not need to compute here. The set $V_{P_1,\ldots,P_d} \subset k^d$ in (8.3) can thus be viewed as the fibre over $(P_1, \ldots, P_d)$ of the algebraic set $V_* \subset W \times k^d$, defined as

$$V_* := \{(P_1, \ldots, P_d, x_1, \ldots, x_d) \in W \times k^d :$$
$$P_1(x_1, \ldots, x_d) = \ldots = P_d(x_1, \ldots, x_d) = 0 \text{ for all } 1 \leq i \leq d\}.$$

This algebraic set is cut out by d polynomial equations on $W \times k^d$ and thus is the union of finitely many algebraic varieties $V_* = V_{*,1} \cup \ldots \cup V_{*,r}$, each of which has codimension at most d in $W \times k^d$. In particular, $\dim(V_{*,i}) \geq \dim(W)$.

Consider one of the components $V_{*,i}$ of V_*. Its projection $W_{*,i}$ in W will be Zariski dense in its closure $\overline{W_{*,i}}$, which is necessarily an irreducible variety since $V_{*,i}$ is, and the projection map from $V_{*,i}$ to $\overline{W_{*,i}}$ will, by definition, be a *dominant map*. By basic algebraic geometry, the preimages in $V_{*,i}$ of a point $(P_1, \ldots, P_d)$ in $W_{*,i}$ will generically have dimension $\dim(V_{*,i}) - \dim(W_{*,i})$ (i.e. this dimension will occur outside of a algebraic subset of $W_{*,i}$ of positive codimension), and even in the non-generic case will have dimension *at least* $\dim(V_{*,i}) - \dim(W_{*,i})$. Since $\dim(V_{*,i}) \geq \dim(W)$, we thus see that the only components $V_{*,i}$ that can contribute to the zero-dimensional set V^0

are those with $\dim(V_{*,i}) = \dim(W_{*,i}) = \dim(W)$. Now, the projection map is a dominant map between two varieties of the same dimension, and is thus generically finite, with the preimages generically having some constant cardinality $D_{*,i}$, and non-generically the preimages have a zero-dimensional component of at most[2] $D_{*,i}$ points.

As a consequence of this analysis, we see that the generic fibre always has at least as many zero-dimensional components as a non-generic fibre, and so to establish Theorem 8.4.5, it suffices to do so for generic $P_1, \ldots, P_m$.

Now take $P_1, \ldots, P_m$ to be generic. We know that generically, the set V is finite; we seek to bound its cardinality $|V|$ by $D_1 \ldots D_m$. To do this, we dualise the problem. Let A be the space of all affine-linear forms $\lambda : k^d \to k$; this is a $d+1$-dimensional vector space. We consider the set $\hat{V}$ of all affine-linear forms λ whose kernel $\{\lambda = 0\}$ intersects V. This is a union of $|V|$ hyperplanes in A, and is thus a hypersurface of degree $|V|$. Thus, to upper bound the size of V, it suffices to upper bound the degree of the hypersurface $\hat{V}$, and this can be done by finding a non-zero polynomial of controlled degree that vanishes identically on $\hat{V}$. The point of this observation is that the property of a polynomial being non-zero is a Zariski-open condition, and so we have a chance of establishing the generic case from a special case.

Now let us look for a (generically non-trivial) polynomial of degree at most $\prod_{i=1}^d D_i$ that vanishes on $\hat{V}$. The idea is to try to dualise the assertion that the monomials $x_1^{a_1} \ldots x_d^{a_d}$ with $a_j < D_j$ for all $1 \leq j \leq d$ generically span the function ring of V, to become a statement about $\hat{V}$.

Let D be a sufficiently large integer (any integer larger than $\sum_{i=1}^d (D_i - 1)$ will do, actually), and let X be the space of all polynomials $P : k^d \to k$ of degree at most D. This is a finite-dimensional vector space over k, generated by the monomials $x_1^{a_1} \ldots x_d^{a_d}$ with $a_1, \ldots, a_d$ non-negative integers adding up to at most D. We can split X as a direct sum

$$X = (\sum_{i=1}^d x_i^{D_i} \cdot X_i) + X_0 \tag{8.4}$$

where for $i, \ldots, d$, X_i is generated by those monomials $x_1^{a_1} \ldots x_d^{a_d}$ of degree at most $D - D_i$ with $a_j < D_j$ for all $j < i$ and with $a_i \geq 0$, and X_0 is generated by those monomials $x_1^{a_1} \ldots x_d^{a_d}$ with $a_j < D_j$ for all $1 \leq j \leq d$. In

[2] Topologically, one can see this claim (at least in the model case $k = \mathbf{C}$) as follows. Any isolated point p in a non-generic preimage must perturb to a zero-dimensional set or to an empty set in nearby preimages, since $V_{*,i}$ is closed. On a generic preimage, p must perturb to a zero-dimensional set (for if it generically perturbs to the empty set, the dimension of $V_{*,i}$ will be too small); since a generic preimage has $D_{*,i}$ points, we conclude that the non-generic preimage can contain at most $D_{*,i}$ isolated points. For a more detailed proof of this claim, see [**He1983**, Proposition 1].

particular,

$$\dim(X) = \sum_{i=1}^{d} \dim(X_i) + \dim(X_0). \tag{8.5}$$

Also observe that

$$\dim(X_0) = \prod_{i=1}^{d} D_i. \tag{8.6}$$

Now let $\lambda \in A$ be an affine-linear form, and consider the sum

$$(\sum_{i=1}^{d} P_i \cdot X_i) + \lambda \cdot X_0. \tag{8.7}$$

This is a subspace of X for D large enough. In view of (8.4), we see that this sum can equal all of X in the case when $P_i = x_i^{D_i}$ and $\lambda = 1$. From (8.5), the property of (8.7) spanning all of X is a Zariski-open condition, and thus we see that (8.7) spans X for generic choices of $P_1, \ldots, P_d$ and λ.

On the other hand, suppose that $\lambda \in \hat{V}$, thus $\lambda(a) = P_1(a) = \ldots = P_d(a) = 0$ for some $a \in k^d$. Then observe that each factor of (8.7) lies in the hyperplane $\{P \in X : P(a) = 0\}$ of X, and so (8.7) does not span X in this case. Thus, for generic $P_1, \ldots, P_d$, we see that (8.7) spans X for generic λ but not for any λ in $\hat{V}$.

The property of (8.7) spanning X is equivalent to a certain resultant-like determinant of a $\dim(X) \times \dim(X)$ matrix being non-zero, where the rows are given by the generators of $P_i \cdot X_i$ and $\lambda \cdot X_0$. For generic $P_1, \ldots, P_d$, this determinant is a non-trivial polynomial in $\lambda \cdot X_0$ which vanishes on $\hat{V}$; an inspection of the matrix reveals that this determinant has degree $\dim(X_0) = \prod_{i=1}^{d} D_i$ in λ. Thus we have found a non-trivial polynomial of degree $\prod_{i=1}^{d} D_i$ that vanishes on $\hat{V}$, and the claim follows.

8.4.2. The overdetermined case. Now we establish Theorem 8.4.6. Fix $P_1, \ldots, P_m$, and let V' be the union of all the positive-dimensional components of V, thus $V_0 = V \backslash V'$.

The main idea here is to perturb $P_1, \ldots, P_m$ to be a *regular sequence* of polynomials outside of V'. More precisely, we have

Lemma 8.4.9 (Regular sequence)**.** *For any $1 \leq r \leq d$, one can find polynomials $Q_1, \ldots, Q_r$, such that for each $i = 1, \ldots, r$, Q_i is a linear combination of $P_i, \ldots, P_m$ (and thus has degree at most D_i), with the P_i coefficient being non-zero, and the set $\{x \in k^d \backslash V' : Q_1 = \ldots = Q_r = 0\}$ in k^d has dimension at most $d - r$ (in the sense that it is covered by finitely many varieties of dimension at most $d - r$).*

Proof. We establish this claim by induction on r. For $r = 1$ the claim follows by setting $Q_1 := P_1$. Now suppose inductively that $1 < r \leq d$, and that $Q_1, \ldots, Q_{r-1}$ have already been found with the stated properties for $r - 1$.

By construction, the polynomials $P_1, \ldots, P_m$ are linear combinations of $Q_1, \ldots, Q_{r-1}, P_r, \ldots, P_m$. Thus, on the set $W := \{x \in k^d \backslash V' : Q_1 = \ldots = Q_{r-1} = 0\}$, the polynomials $P_r, \ldots, P_m$ can only simultaneously vanish on the zero-dimensional set $V \backslash V' = V_0$. On the other hand, each irreducible component of W has dimension $d - r + 1$, which is positive. Thus it is not possible for $P_r, \ldots, P_m$ to simultaneously vanish on any of these components. If one then sets Q_r to be a generic linear combination of $P_r, \ldots, P_m$, then we thus see that Q_r will also not vanish on any of these components, and so $\{x \in k^d \backslash V' : Q_1 = \ldots = Q_r = 0\}$ has dimension at most $d - r$. Also, generically the P_r coefficient of Q_r is non-zero, and the claim follows. □

From the above lemma (with $r := d$) we see that V_0 is contained in the set $\{Q_1 = \ldots = Q_d = 0\}$. By Theorem 8.4.5, the latter set has cardinality at most $D_1 \ldots D_d$, and the claim follows.

8.5. The Brunn-Minkowski inequality in nilpotent groups

One of the fundamental inequalities in convex geometry is the *Brunn-Minkowski inequality*, which asserts that if A, B are two non-empty bounded open subsets of $\mathbf{R}^d$, then

$$\mu(A + B)^{1/d} \geq \mu(A)^{1/d} + \mu(B)^{1/d}, \tag{8.8}$$

where

$$A + B := \{a + b : a \in A, b \in B\}$$

is the sumset of A and B, and μ denotes Lebesgue measure. The estimate is sharp, as can be seen by considering the case when A, B are convex bodies that are dilates of each other, thus $A = \lambda B := \{\lambda b : b \in B\}$ for some $\lambda > 0$, since in this case one has $\mu(A) = \lambda^d \mu(B)$, $A + B = (\lambda + 1)B$, and $\mu(A + B) = (\lambda + 1)^d \mu(B)$.

The Brunn-Minkowski inequality has many applications in convex geometry. To give just one example, if we assume that A has a smooth boundary ∂A, and set B equal to a small ball $B = B(0, \varepsilon)$, then $\mu(B)^{1/d} = \varepsilon \mu(B(0, 1))^{1/d}$, and in the limit $\varepsilon \to 0$ one has

$$\mu(A + B) = \mu(A) + \varepsilon |\partial A| + o(\varepsilon)$$

where $|\partial A|$ is the surface measure of A; applying the Brunn-Minkowski inequality and performing a Taylor expansion, one soon arrives at the *isoperimetric inequality*

$$|\partial A| \geq d \mu(A)^{1-1/d} \mu(B(0, 1))^{1/d}.$$

Thus one can view the isoperimetric inequality as an infinitesimal limit of the Brunn-Minkowski inequality.

There are many proofs known of the Brunn-Minkowski inequality. Firstly, the inequality is trivial in one dimension:

Lemma 8.5.1 (One-dimensional Brunn-Minkowski)**.** *If $A, B \subset \mathbf{R}$ are non-empty measurable sets, then*

$$\mu(A+B) \geq \mu(A) + \mu(B).$$

Proof. By *inner regularity* we may assume that A, B are compact. The claim then follows since $A+B$ contains the sets $\sup(A)+B$ and $A+\inf(B)$, which meet only at a single point $\sup(A)+\inf(B)$. □

For the higher dimensional case, the inequality can be established from the *Prékopa-Leindler inequality*:

Theorem 8.5.2 (Prékopa-Leindler inequality in $\mathbf{R}^d$)**.** *Let $0 < \theta < 1$, and let $f, g, h : \mathbf{R}^d \to \mathbf{R}$ be non-negative measurable functions obeying the inequality*

$$h(x+y) \geq f(x)^{1-\theta} g(y)^\theta \tag{8.9}$$

for all $x, y \in \mathbf{R}^d$. Then we have

$$\int_{\mathbf{R}^d} h \geq \frac{1}{(1-\theta)^{d(1-\theta)}\theta^{d\theta}} (\int_{\mathbf{R}^d} f)^{1-\theta} (\int_{\mathbf{R}^d} g)^\theta. \tag{8.10}$$

This inequality is usually stated using $h((1-\theta)x+\theta y)$ instead of $h(x+y)$ in order to eliminate the ungainly factor $\frac{1}{(1-\theta)^{d(1-\theta)}\theta^{d\theta}}$. However, we formulate the inequality in this fashion in order to avoid any reference to the dilation maps $x \mapsto \lambda x$; the reason for this will become clearer later.

The Prékopa-Leindler inequality quickly implies the Brunn-Minkowski inequality. Indeed, if we apply it to the indicator functions $f := 1_A, g := 1_B, h := 1_{A+B}$ (which certainly obey (8.9)), then (8.10) gives

$$\mu(A+B)^{1/d} \geq \frac{1}{(1-\theta)^{1-\theta}\theta^\theta} \mu(A)^{\frac{1-\theta}{d}} \mu(B)^{\frac{\theta}{d}}$$

for any $0 < \theta < 1$. We can now optimise in θ; the optimal value turns out to be

$$\theta := \frac{\mu(B)^{1/d}}{\mu(A)^{1/d} + \mu(B)^{1/d}}$$

which yields (8.8).

To prove the Prékopa-Leindler inequality, we first observe that the inequality *tensorises* in the sense that if it is true in dimensions d_1 and d_2, then it is automatically true in dimension d_1+d_2. Indeed, if $f, g, h : \mathbf{R}^{d_1} \times \mathbf{R}^{d_2} \to \mathbf{R}^+$ are measurable functions obeying (8.9) in dimension d_1+d_2, then for any $x_1, y_1 \in \mathbf{R}^{d_1}$, the functions $f(x_1, \cdot), g(y_1, \cdot), h(x_1+y_1, \cdot) : \mathbf{R}^{d_2} \to \mathbf{R}^+$

obey (8.9) in dimension d_2. Applying the Prékopa-Leindler inequality in dimension d_2, we conclude that

$$H(x_1 + y_1) \geq \frac{1}{(1-\theta)^{d_2(1-\theta)}\theta^{d_2\theta}} F(x_1)^{1-\theta} G(y_1)^\theta$$

for all $x_1, y_1 \in \mathbf{R}^{d_1}$, where $F(x_1) := \int_{\mathbf{R}^{d_2}} f(x_1, x_2)\ dx_2$ and similarly for G, H. But then if we apply the Prékopa-Leindler inequality again, this time in dimension d_1 and to the functions F, G, and $(1-\theta)^{d_2(1-\theta)}\theta^{d_2\theta}H$, and then use the *Fubini-Tonelli theorem*, we obtain (8.10).

From tensorisation, we see that to prove the Prékopa-Leindler inequality it suffices to do so in the one-dimensional case. We can derive this from Lemma 8.5.1 by reversing the "Prékopa-Leindler implies Brunn-Minkowski" argument given earlier, as follows. If (8.9) holds (in one dimension), then the super-level sets $\{f > \lambda\}, \{g > \lambda\}, \{h > \lambda\}$ are related by the set-theoretic inclusion

$$\{h > \lambda\} \supset \{f > \lambda\} + \{g > \lambda\}$$

and thus by Lemma 8.5.1

$$\mu(\{h > \lambda\}) \geq \mu(\{f > \lambda\}) + \mu(\{g > \lambda\}).$$

On the other hand, from the Fubini-Tonelli theorem one has the distributional identity

$$\int_{\mathbf{R}} h = \int_0^\infty \mu(\{h > \lambda\})\ d\lambda$$

(and similarly for f, g), and thus

$$\int_{\mathbf{R}} h \geq \int_{\mathbf{R}} f + \int_{\mathbf{R}} g.$$

The claim then follows from the *weighted arithmetic mean-geometric mean inequality* $(1-\theta)x + \theta y \geq x^{1-\theta} y^\theta$.

In this section we will make the simple observation (which appears in [**LeMa2005**] in the case of the Heisenberg group, but may have also been stated elsewhere in the literature) that the above argument carries through without much difficulty to the nilpotent setting, to give a nilpotent Brunn-Minkowski inequality:

Theorem 8.5.3 (Nilpotent Brunn-Minkowski)**.** *Let G be a connected, simply connected nilpotent Lie group of (topological) dimension d, and let A, B be bounded open subsets of G. Let μ be a* Haar measure *on G (note that nilpotent groups are* unimodular*, so there is no distinction between left and right Haar measure). Then*

$$\mu(A \cdot B)^{1/d} \geq \mu(A)^{1/d} + \mu(B)^{1/d}. \tag{8.11}$$

Here of course $A \cdot B := \{ab : a \in A, b \in B\}$ is the product set of A and B.

Indeed, by repeating the previous arguments, the nilpotent Brunn-Minkowski inequality will follow from

Theorem 8.5.4 (Nilpotent Prékopa-Leindler inequality)**.** *Let G be a connected, simply connected nilpotent Lie group of topological dimension d with a Haar measure μ. Let $0 < \theta < 1$, and let $f, g, h : G \to \mathbf{R}$ be non-negative measurable functions obeying the inequality*

$$h(xy) \geq f(x)^{1-\theta} g(y)^{\theta} \tag{8.12}$$

for all $x, y \in G$. Then we have

$$\int_G h \, d\mu \geq \frac{1}{(1-\theta)^{d(1-\theta)} \theta^{d\theta}} (\int_G f \, d\mu)^{1-\theta} (\int_G g \, d\mu)^{\theta}. \tag{8.13}$$

To prove the nilpotent Prékopa-Leindler inequality, the key observation is that this inequality not only tensorises; it *splits* with respect to short exact sequences. Indeed, suppose one has a short exact sequence

$$0 \to K \to G \to H \to 0$$

of connected, simply connected nilpotent Lie groups. The adjoint action of the connected group G on K acts nilpotently on the Lie algebra of K and is thus unimodular. Because of this, we can split a Haar measure μ_G on G into Haar measures μ_K, μ_H on K, H respectively so that we have the Fubini-Tonelli formula

$$\int_G f(g) \, d\mu_G(g) = \int_H F(h) \, d\mu_H(h)$$

for any measurable $f : G \to \mathbf{R}^+$, where $F(h)$ is defined by the formula

$$F(h) := \int_K f(kg) d\mu_K(k) = \int_K f(gk) \, d\mu_K(k)$$

for any coset representative $g \in G$ of h (the choice of g is not important, thanks to unimodularity of the conjugation action). It is then not difficult to repeat the proof of tensorisation (relying heavily on the unimodularity of conjugation) to conclude that the nilpotent Prékopa-Leindler inequality for H and K implies the Prékopa-Leindler inequality for G; we leave this as an exercise to the interested reader.

Now if G is a connected simply connected Lie group, then the abelianisation $G/[G,G]$ is connected and simply connected and thus isomorphic to a vector space. This implies that $[G,G]$ is a retract of G and is thus also connected and simply connected. From this and an induction of the step of the nilpotent group, we see that the nilpotent Prékopa-Leindler inequality

follows from the abelian case, which we have already established in Theorem 8.5.2.

Remark 8.5.5. Some connected, simply connected nilpotent groups G (and specifically, the *Carnot groups*) can be equipped with a one-parameter family of dilations $x \mapsto \lambda \cdot x$, which are a family of automorphisms on G, which dilate the Haar measure by the formula

$$\mu(\lambda \cdot x) = \lambda^D \mu(x)$$

for an integer D, called the *homogeneous dimension* of G, which is typically larger than the topological dimension. For instance, in the case of the Heisenberg group

$$G := \begin{pmatrix} 1 & \mathbf{R} & \mathbf{R} \\ 0 & 1 & \mathbf{R} \\ 0 & 0 & 1 \end{pmatrix},$$

which has topological dimension $d = 3$, the natural family of dilations is given by

$$\lambda : \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \mapsto \begin{pmatrix} 1 & \lambda x & \lambda^2 z \\ 0 & 1 & \lambda y \\ 0 & 0 & 1 \end{pmatrix}$$

with homogeneous dimension $D = 4$. Because the two notions d, D of dimension are usually distinct in the nilpotent case, it is no longer helpful to try to use these dilations to simplify the proof of the Brunn-Minkowski inequality, in contrast to the Euclidean case. This is why we avoided using dilations in the preceding discussion. It is natural to wonder whether one could replace d by D in (8.11), but it can be easily shown that the exponent d is best possible (an observation that essentially appeared first in [**Mo2003**]). Indeed, working in the Heisenberg group for sake of concreteness, consider the set

$$A := \{ \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} : |x|, |y| \leq N, |z| \leq N^{10} \}$$

for some large parameter N. This set has measure N^{12} using the standard Haar measure on G. The product set $A \cdot A$ is contained in

$$A := \{ \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} : |x|, |y| \leq 2N, |z| \leq 2N^{10} + O(N^2) \}$$

and thus has measure at most $8N^{12} + O(N^4)$. This already shows that the exponent in (8.11) cannot be improved beyond $d = 3$; note that the homogeneous dimension $D = 4$ is making its presence known in the $O(N^4)$ term in the measure of $A \cdot A$, but this is a lower order term only.

It is somewhat unfortunate that the nilpotent Brunn-Minkowski inequality is adapted to the topological dimension rather than the homogeneous one, because it means that some of the applications of the inequality (such as the application to isoperimetric inequalities mentioned at the start of the section) break down[3].

Remark 8.5.6. The inequality can be extended to non-simply-connected connected nilpotent groups G, if d is now set to the dimension of the largest simply connected quotient of G. It seems to me that this is the best one can do in general; for instance, if G is a torus, then the inequality fails for any $d > 0$, as can be seen by setting $A = B = G$.

Remark 8.5.7. Specialising the nilpotent Brunn-Minkowski inequality to the case $A = B$, we conclude that

$$\mu(A \cdot A) \geq 2^d \mu(A).$$

This inequality actually has a much simpler proof (attributed to Tsachik Gelander in [**Hr2012**]): one can show that for a connected, simply connected Lie group G, the exponential map $\exp : \mathfrak{g} \to G$ is a measure-preserving homeomorphism, for some choice of Haar measure $\mu_{\mathfrak{g}}$ on $\mathfrak{g}$, so it suffices to show that

$$\mu_{\mathfrak{g}}(\log(A \cdot A)) \geq 2^d \mu_{\mathfrak{g}}(\log A).$$

But $A \cdot A$ contains all the squares $\{a^2 : a \in A\}$ of A, so $\log(A \cdot A)$ contains the isotropic dilation $2 \cdot \log A$, and the claim follows. Note that if we set A to be a small ball around the origin, we can modify this argument to give another demonstration of why the topological dimension d cannot be replaced with any larger exponent in (8.11).

One may tentatively conjecture that the inequality $\mu(A \cdot A) \geq 2^d \mu(A)$ in fact holds in all unimodular connected, simply connected Lie groups G, and all bounded open subsets A of G; I do not know if this bound is always true, however.

[3]Indeed, the topic of isoperimetric inequalities for the Heisenberg group is a subtle one, with many naive formulations of the inequality being false. See [**Mo2003**] for more discussion.

Chapter 9

Dynamics

9.1. The Furstenberg recurrence theorem and finite extensions

In [**Fu1977**], Furstenberg established his celebrated multiple recurrence theorem:

Theorem 9.1.1 (Furstenberg multiple recurrence)**.** *Let $(X, \mathcal{B}, \mu, T)$ be a measure-preserving system, thus $(X, \mathcal{B}, \mu)$ is a probability space and $T : X \to X$ is a measure-preserving bijection such that T and T^{-1} are both measurable. Let E be a measurable subset of X of positive measure $\mu(E) > 0$. Then for any $k \geq 1$, there exists $n > 0$ such that*

$$E \cap T^{-n}E \cap \ldots \cap T^{-(k-1)n}E \neq \emptyset.$$

Equivalently, there exists $n > 0$ and $x \in X$ such that

$$x, T^n x, \ldots, T^{(k-1)n}x \in E.$$

As is well known, the Furstenberg multiple recurrence theorem is equivalent to Szemerédi's theorem [**Sz1975**], thanks to the Furstenberg correspondence principle; see for instance [**Ta2009**, §2.10].

The multiple recurrence theorem is proven, roughly speaking, by an induction on the "complexity" of the system $(X, \mathcal{X}, \mu, T)$. Indeed, for very simple systems, such as periodic systems (in which T^n is the identity for some $n > 0$, which is for instance the case for the circle shift $X = \mathbf{R}/\mathbf{Z}$, $Tx := x+\alpha$ with a rational shift α), the theorem is trivial; at a slightly more advanced level, *almost periodic* (or *compact*) systems (in which $\{T^n f : n \in \mathbf{Z}\}$ is a precompact subset of $L^2(X)$ for every $f \in L^2(X)$, which is for instance the case for irrational circle shifts), is also quite easy. One then shows

that the multiple recurrence property is preserved under various *extension* operations (specifically, compact extensions, weakly mixing extensions, and limits of chains of extensions), which then gives the multiple recurrence theorem as a consequence of the *Furstenberg-Zimmer structure theorem* for measure-preserving systems. See [**Ta2009**, §2.15] for further discussion.

From a high-level perspective, this is still one of the most conceptual proofs known of Szemerédi's theorem. However, the individual components of the proof are still somewhat intricate. Perhaps the most difficult step is the demonstration that the multiple recurrence property is preserved under *compact extensions*; see for instance [**Ta2009**, §2.13], which is devoted entirely to this step. This step requires quite a bit of measure-theoretic and/or functional analytic machinery, such as the theory of disintegrations, relatively almost periodic functions, or Hilbert modules.

However, I recently realised that there is a special case of the compact extension step - namely that of *finite* extensions - which avoids almost all of these technical issues while still capturing the essence of the argument (and in particular, the key idea of using *van der Waerden's theorem* [**vdW1927**]). As such, this may serve as a pedagogical device for motivating this step of the proof of the multiple recurrence theorem.

Let us first explain what a finite extension is. Given a measure-preserving system $X = (X, \mathcal{X}, \mu, T)$, a finite set Y, and a measurable map $\rho : X \to \mathrm{Sym}(Y)$ from X to the permutation group of Y, one can form the *finite extension*

$$X \ltimes_\rho Y = (X \times Y, \mathcal{X} \times \mathcal{Y}, \mu \times \nu, S),$$

which as a probability space is the product of $(X, \mathcal{X}, \mu)$ with the finite probability space $Y = (Y, \mathcal{Y}, \nu)$ (with the discrete σ-algebra and uniform probability measure), and with shift map

$$S(x, y) := (Tx, \rho(x)y). \tag{9.1}$$

One easily verifies that this is indeed a measure-preserving system. We refer to ρ as the *cocycle* of the system.

An example of finite extensions comes from group theory. Suppose we have a short exact sequence

$$0 \to K \to G \to H \to 0$$

of finite groups. Let g be a group element of G, and let h be its projection in H. Then the shift map $x \mapsto gx$ on G (with the discrete σ-algebra and uniform probability measure) can be viewed as a finite extension of the shift map $y \mapsto hy$ on H (again with the discrete σ-algebra and uniform probability measure), by arbitrarily selecting a section $\phi : H \to G$ that inverts the projection map, identifying G with $H \times K$ by identifying $k\phi(y)$

with (y,k) for $y \in H, k \in K$, and using the cocycle

$$\rho(y) := \phi(hy)^{-1} g \phi(y).$$

Thus, for instance, the unit shift $x \mapsto x+1$ on $\mathbf{Z}/N\mathbf{Z}$ can be thought of as a finite extension of the unit shift $x \mapsto x+1$ on $\mathbf{Z}/M\mathbf{Z}$ whenever N is a multiple of M.

Another example comes from Riemannian geometry. If M is a Riemannian manifold that is a finite cover of another Riemannian manifold N (with the metric on M being the pullback of that on N), then (unit time) geodesic flow on the cosphere bundle of M is a finite extension of the corresponding flow on N.

Here, then, is the finite extension special case of the compact extension step in the proof of the multiple recurrence theorem:

Proposition 9.1.2 (Finite extensions)**.** *Let $X \rtimes_\rho Y$ be a finite extension of a measure-preserving system X. If X obeys the conclusion of the Furstenberg multiple recurrence theorem, then so does $X \rtimes_\rho Y$.*

Before we prove this proposition, let us first give the combinatorial analogue.

Lemma 9.1.3. *Let A be a subset of the integers that contains arbitrarily long arithmetic progressions, and let $A = A_1 \cup \ldots \cup A_M$ be a colouring of A by M colours (or equivalently, a partition of A into M colour classes A_i). Then at least one of the A_i contains arbitrarily long arithmetic progressions.*

Proof. By the infinite pigeonhole principle, it suffices to show that for each $k \geq 1$, one of the colour classes A_i contains an arithmetic progression of length k.

Let N be a large integer (depending on k and M) to be chosen later. Then A contains an arithmetic progression of length N, which may be identified with $\{0, \ldots, N-1\}$. The colouring of A then induces a colouring on $\{0, \ldots, N-1\}$ into M colour classes. Applying (the finitary form of) *van der Waerden's theorem* [**vdW1927**], we conclude that if N is sufficiently large depending on M and k, then one of these colouring classes contains an arithmetic progression of length k; undoing the identification, we conclude that one of the A_i contains an arithmetic progression of length k, as desired. □

Of course, by specialising to the case $A = \mathbf{Z}$, we see that the above Lemma is in fact equivalent to van der Waerden's theorem.

Now we prove Proposition 9.1.2.

Proof. Fix k. Let E be a positive measure subset of $X \rtimes_\rho Y = (X \times Y, \mathcal{X} \times \mathcal{Y}, \mu \times \nu, S)$. By Fubini's theorem, we have

$$\mu \times \nu(E) = \int_X f(x)\ d\mu(x)$$

where $f(x) := \nu(E_x)$ and $E_x := \{y \in Y : (x,y) \in E\}$ is the fibre of E at x. Since $\mu \times \nu(E)$ is positive, we conclude that the set

$$F := \{x \in X : f(x) > 0\} = \{x \in X : E_x \neq \emptyset\}$$

is a positive measure subset of X. Note for each $x \in F$, we can find an element $g(x) \in Y$ such that $(x, g(x)) \in E$. While not strictly necessary for this argument, one can ensure if one wishes that the function g is measurable by totally ordering Y, and then letting $g(x)$ the minimal element of Y for which $(x, g(x)) \in E$.

Let N be a large integer (which will depend on k and the cardinality M of Y) to be chosen later. Because X obeys the multiple recurrence theorem, we can find a positive integer n and $x \in X$ such that

$$x, T^n x, T^{2n} x, \ldots, T^{(N-1)n} x \in F.$$

Now consider the sequence of N points

$$S^{-mn}(T^{mn}x, g(T^{mn}x))$$

for $m = 0, \ldots, N-1$. From (9.1), we see that

$$S^{-mn}(T^{mn}x, g(T^{mn}x)) = (x, c(m)) \tag{9.2}$$

for some sequence $c(0), \ldots, c(N-1) \in Y$. This can be viewed as a colouring of $\{0, \ldots, N-1\}$ by M colours, where M is the cardinality of Y. Applying *van der Waerden's theorem*, we conclude (if N is sufficiently large depending on k and $|Y|$) that there is an arithmetic progression $a, a+r, \ldots, a+(k-1)r$ in $\{0, \ldots, N-1\}$ with $r > 0$ such that

$$c(a) = c(a+r) = \ldots = c(a+(k-1)r) = c$$

for some $c \in Y$. If we then let $y = (x, c)$, we see from (9.2) that

$$S^{an+irn} y = (T^{(a+ir)n} x, g(T^{(a+ir)n} x)) \in E$$

for all $i = 0, \ldots, k-1$, and the claim follows. □

Remark 9.1.4. The precise connection between Lemma 9.1.3 and Proposition 9.1.2 arises from the following observation: with E, F, g as in the proof of Proposition 9.1.2, and $x \in X$, the set

$$A := \{n \in \mathbf{Z} : T^n x \in F\}$$

can be partitioned into the classes

$$A_i := \{n \in \mathbf{Z} : S^n(x,i) \in E'\}$$

where $E' := \{(x, g(x)) : x \in F\} \subset E$ is the graph of g. The multiple recurrence property for X ensures that A contains arbitrarily long arithmetic progressions, and so therefore one of the A_i must also, which gives the multiple recurrence property for Y.

Remark 9.1.5. Compact extensions can be viewed as a generalisation of finite extensions, in which the fibres are no longer finite sets, but are themselves measure spaces obeying an additional property, which roughly speaking asserts that for many functions f on the extension, the shifts $T^n f$ of f behave in an almost periodic fashion on most fibres, so that the orbits $T^n f$ become totally bounded on each fibre. This total boundedness allows one to obtain an analogue of the above colouring map $c()$ to which van der Waerden's theorem can be applied.

9.2. Rohlin's problem

Let $G = (G, +)$ be an abelian countable discrete group. A measure-preserving G-system $X = (X, \mathcal{X}, \mu, (T_g)_{g\in G})$ (or *G-system for short*) is a probability space $(X, \mathcal{X}, \mu)$, equipped with a measure-preserving action $T_g : X \to X$ of the group G, thus

$$\mu(T_g(E)) = \mu(E)$$

for all $E \in \mathcal{X}$ and $g \in G$, and

$$T_g T_h = T_{g+h}$$

for all $g, h \in G$, with T_0 equal to the identity map. Classically, ergodic theory has focused on the cyclic case $G = \mathbf{Z}$ (in which the T_g are iterates of a single map $T = T_1$, with elements of G being interpreted as a time parameter), but one can certainly consider actions of other groups G also (including continuous or non-abelian groups).

A G-system is said to be *strongly 2-mixing*, or *strongly mixing* for short, if one has

$$\lim_{g\to\infty} \mu(A \cap T_g B) = \mu(A)\mu(B)$$

for all $A, B \in \mathcal{X}$, where the convergence is with respect to the one-point compactification of G (thus, for every $\varepsilon > 0$, there exists a compact (hence finite) subset K of G such that $|\mu(A \cap T_g B) - \mu(A)\mu(B)| \leq \varepsilon$ for all $g \notin K$).

Similarly, we say that a G-system is *strongly 3-mixing* if one has

$$\lim_{g,h,h-g\to\infty} \mu(A \cap T_g B \cap T_h C) = \mu(A)\mu(B)\mu(C)$$

for all $A, B, C \in \mathcal{X}$, thus for every $\varepsilon > 0$, there exists a finite subset K of G such that

$$|\mu(A \cap T_g B \cap T_h C) - \mu(A)\mu(B)\mu(C)| \leq \varepsilon$$

whenever $g, h, h - g$ all lie outside K.

It is obvious that a strongly 3-mixing system is necessarily strong 2-mixing. In the case of **Z**-systems, it has been an open problem for some time, due to Rohlin [**Ro1949**], whether the converse is true:

Problem 9.2.1 (Rohlin's problem)**.** Is every strongly mixing **Z**-system necessarily strongly 3-mixing?

This is a surprisingly difficult problem. In the positive direction, a routine application of the Cauchy-Schwarz inequality (via van der Corput's inequality) shows that every strongly mixing system is *weakly* 3*-mixing*, which roughly speaking means that $\mu(A \cap T_g B \cap T_h C)$ converges to $\mu(A)\mu(B)\mu(C)$ for *most* $g, h \in \mathbf{Z}$. Indeed, every weakly mixing system is in fact weakly mixing of all orders; see for instance [**Ta2009**, §2.10]. So the problem is to exclude the possibility of correlation between A, $T_g B$, and $T_h C$ for a small but non-trivial number of pairs (g, h).

It is also known that the answer to Rohlin's problem is affirmative for rank one transformations [**Ka1984**] and for shifts with purely singular continuous spectrum [**Ho1991**] (note that strongly mixing systems cannot have any non-trivial point spectrum). Indeed, any counterexample to the problem, if it exists, is likely to be highly pathological.

In the other direction, Rohlin's problem is known to have a negative answer for $\mathbf{Z}^2$-systems, by a well-known counterexample of Ledrappier [**Le1978**] which can be described as follows. One can view a $\mathbf{Z}^2$-system as being essentially equivalent to a *stationary process* $(x_{n,m})_{(n,m)\in\mathbf{Z}^2}$ of random variables $x_{n,m}$ in some range space Ω indexed by $\mathbf{Z}^2$, with X being $\Omega^{\mathbf{Z}^2}$ with the obvious shift map

$$T_{(g,h)}(x_{n,m})_{(n,m)\in\mathbf{Z}^2} := (x_{n-g,m-h})_{(n,m)\in\mathbf{Z}^2}.$$

In Ledrappier's example, the $x_{n,m}$ take values in the finite field $\mathbf{F}_2$ of two elements, and are selected at uniformly random subject to the "Pascal's triangle" linear constraints

$$x_{n,m} = x_{n-1,m} + x_{n,m-1}.$$

A routine application of the *Kolmogorov extension theorem* (see e.g. [**Ta2011**, §1.7]) allows one to build such a process. The point is that due to the properties of Pascal's triangle modulo 2 (known as *Sierpinski's triangle*), one has

$$x_{n,m} = x_{n-2^k,m} + x_{n,m-2^k}$$

for all powers of two 2^k. This is enough to destroy strong 3-mixing, because it shows a strong correlation between x, $T_{(2^k,0)}x$, and $T_{(0,2^k)}x$ for arbitrarily large k and randomly chosen $x \in X$. On the other hand, one can still show that x and $T_g x$ are asymptotically uncorrelated for large g, giving strong 2-mixing. Unfortunately, there are significant obstructions to converting

Ledrappier's example from a $\mathbf{Z}^2$-system to a $\mathbf{Z}$-system, as pointed out in [**de2006**].

In this section, I would like to record a "finite field" variant of Ledrappier's construction, in which $\mathbf{Z}^2$ is replaced by the function field ring $\mathbf{F}_3[t]$, which is a "dyadic" (or more precisely, "triadic") model for the integers (cf. [**Ta2008**, §1.6]). In other words:

Theorem 9.2.2. *There exists a $\mathbf{F}_3[t]$-system that is strongly 2-mixing but not strongly 3-mixing.*

The idea is much the same as that of Ledrappier; one builds a stationary $\mathbf{F}_3[t]$-process $(x_n)_{n\in\mathbf{F}_3[t]}$ in which $x_n \in \mathbf{F}_3$ are chosen uniformly at random subject to the constraints

$$x_n + x_{n+t^k} + x_{n+2t^k} = 0 \tag{9.3}$$

for all $n \in \mathbf{F}_3[t]$ and all $k \geq 0$. Again, this system is manifestly not strongly 3-mixing, but can be shown to be strongly 2-mixing; I give details below the fold.

As I discussed in [**Ta2008**, §1.6], in many cases the dyadic model serves as a good guide for the non-dyadic model. However, in this case there is a curious rigidity phenomenon that seems to prevent Ledrappier-type examples from being transferable to the one-dimensional non-dyadic setting; once one restores the Archimedean nature of the underlying group, the constraints (9.3) not only reinforce each other strongly, but also force so much linearity on the system that one loses the strong mixing property.

9.2.1. The example. Let B be any ball in $\mathbf{F}_3[t]$, i.e. any set of the form $\{n \in \mathbf{F}_3[t] : \deg(n-n_0) \leq K\}$ for some $n_0 \in \mathbf{F}_3[t]$ and $K \geq 0$. One can then create a process $x_B = (x_n)_{n\in B}$ adapted to this ball, by declaring $(x_n)_{n\in B}$ to be uniformly distributed in the vector space $V_B \leq \mathbf{F}_3^B$ of all tuples with coefficients in $\mathbf{F}_3$ that obey (9.3) for all $n \in B$ and $k \leq K$. Because any translate of a line $(n, n+t^k, n+t^{2k})$ is still a line, we see that this process is stationary with respect to all shifts $n \mapsto n+g$ of degree $\deg(g)$ at most K. Also, if $B \subset B'$ are nested balls, we see that the vector space $V_{B'}$ projects surjectively via the restriction map to V_B (since any tuple obeying (9.3) in B can be extended periodically to one obeying (9.3) in B'). As such, we see that the process x_B is equivalent in distribution to the restriction $x_{B'} \restriction_B$ of $x_{B'}$ to B. Applying *the Kolmogorov extension theorem*, we conclude that there exists an infinite process $x = (x_n)_{n\in\mathbf{F}_3[t]}$ whose restriction $x \restriction_B$ to any ball B has the distribution of x_B. As each x_B was stationary with respect to translations that preserved B, we see that the full process x is stationary with respect to the entire group $\mathbf{F}_3[t]$.

Now let B be a ball

$$B := \{n \in \mathbf{F}_3[t] : \deg(n - n_0) \leq K\},$$

which we divide into three equally sized sub-balls B_0, B_1, B_2 by the formula

$$B_i := \{n \in \mathbf{F}_3[t] : \deg(n - (n_0 + it^K)) \leq K - 1\}.$$

By construction, we see that

$$V_B = \{(x_{B_0}, x_{B_1}, x_{B_2}) : x_{B_0}, x_{B_1}, x_{B_2} \in V_{B_0}; x_{B_0} + x_{B_1} + x_{B_2} = 0\}$$

where we use translation by t^K to identify V_{B_0}, V_{B_1}, and V_{B_2} together. As a consequence, we see that the projection map $(x_{B_0}, x_{B_1}, x_{B_2}) \to (x_{B_0}, x_{B_1})$ from V_B to $V_{B_0} \times V_{B_0}$ is surjective, and this implies that the random variables $x \restriction_{B_0}, x \restriction_{B_1}$ are independent. More generally, this argument implies that for any disjoint balls B, B', the random variables $x \restriction_B$ and $x \restriction_{B'}$ are independent.

Now we can prove strong 2-mixing. Given any measurable event A and any $\varepsilon > 0$, one can find a ball B and a set A' depending only on $x \restriction_B$ such that A and A' differ by at most ε in measure. On the other hand, for g outside of $B - B$, A' and $T_g A'$ are determined by the restrictions of x to disjoint balls and are thus independent. In particular,

$$\mu(A' \cap T_g A') = \mu(A')^2$$

and thus

$$\mu(A \cap T_g A) = \mu(A)^2 + O(\varepsilon)$$

which gives strong 2-mixing.

On the other hand, we have $x_0 + x_{t^k} + x_{2t^k} = 0$ almost surely, while each x_0, x_{t^k}, x_{2t^k} are uniformly distributed in $\mathbf{F}_3$ and pairwise independent. In particular, if E is the event that $x_0 = 0$, we see that

$$\mu(E) = 1/3$$

and

$$\mu(E \cap T_{t^k} E \cap T_{t^{2k}} E) = 1/9$$

showing that strong 3-mixing fails.

Remark 9.2.3. In the Archimedean case $G = \mathbf{Z}$, a constraint such as $x_n + x_{n+1} + x_{n+2} = 0$ propagates itself to force complete linearity of x_n, which is highly incompatible with strong mixing; in contrast, in the non-Archimedean case $G = \mathbf{F}_3$, such a constraint does not propagate very far. It is then tempting to relax this constraint, for instance by adopting an *Ising-type model* which penalises a configuration whenever quantities such as $x_n + x_{n+1} + x_{n+2}$ deviates from zero. However, to destroy strong 3-mixing, one needs infinitely many such penalisation terms, which roughly corresponds to an Ising model in an infinite-dimensional lattice. In such models, it seems

difficult to find a way to set the "temperature" parameters in such a way that one has meaningful 3-correlations, without the system "freezing up" so much that 2-mixing fails. It is also tempting to try to truncate the constraints such as (9.3) to prevent their propagation, but it seems that any naive attempt to perform a truncation either breaks stationarity, or introduces enough periodicity into the system that 2-mixing breaks down. My tentative opinion on this problem is that a **Z**-counterexample is constructible, but one would have to use a very delicate and finely tuned construction to achieve it.

Chapter 10

Miscellaneous

10.1. Worst movie polls

Every so often, one sees on the web some poll for the "worst X", where X is some form of popular entertainment; let's take X to be "movies" for sake of discussion.

Invariably, the results of these polls are somewhat disappointing; a "worst movie list" will often contain examples of bad movies, but with an arbitrary-seeming ranking, with many obviously bad movies missing from the list.

Of course, much of this can be ascribed to the highly subjective and variable nature of the tastes of those being polled, as well as the over-marketing of various mediocre but not exceptionally terrible movies. However, it turns out that even in an idealised situation in which all movie watchers use the same objective standard to rate movies, and where the success of each movie is determined solely by its quality, a worst movie poll will still often give totally inaccurate results.

Informally, the reason for this is that the truly bad movies, by their nature, are so unpopular that most people will not have watched them, and so they rarely even show up on the polls at all.

One can mathematically model this as follows. Let us say there are N movies, ranked in order of highest quality to least. Suppose that the k^{th} best movie has been watched by a proportion p_k of the population. As we are assuming that movie success is determined by quality, we suppose that the p_k are decreasing in k. A randomly selected member of the population thus has a probability p_k of seeing the k^{th} movie. In order to make the analysis tractable, we make the (unrealistic) assumption that these events of seeing the k^{th} movie are independent in k.

As such, the probability that a given voter will rank movie k as the worst movie (because he or she has seen that movie, but has not seen any worse movie) is

$$p_k(1-p_{k+1})\dots(1-p_N). \tag{10.1}$$

The winner of the poll should then be the movie which maximises the quantity (10.1).

One can solve this optimisation problem by assuming a power law

$$p_k \sim ck^{-\alpha}$$

for some parameters c and α, which typically are comparable to 1. It is an instructive exercise to optimise (10.1) using this law. What one finds is

that the value of the exponent α becomes key. If $\alpha < 1$ (and N is large), then (10.1) is maximised at $k = N$, and so in this case the poll should indeed rate the very worst movies at the top of their ranking.

If $\alpha > 1$, there is a surprising reversal; (10.1) is instead maximised for a value of k which is bounded, $k = O(1)$. Basically, the poll now ranks the worst blockbuster movie, rather than the worst movie period; a mediocre but widely viewed movie will beat out a terrible but obscure movie.

Amusingly, according to *Zipf's law*, one expects α to be close to 1. As such, there is a critical phase transition (especially if the constant c is also at the critical value of 1) and now one can anticipate the poll to more or less randomly select movies of any level of quality. So one can blame Zipf's law for the inaccuracy of "worst movie" polls.

10.2. Descriptive and prescriptive science

Broadly speaking, work in an academic discipline can be divided into *descriptive*[1] activity, which seeks to objectively describe the world we live in, and *prescriptive* activity, which is more subjective and seeks to define how the world ought to be interpreted.

However, the division between descriptive and prescriptive activity varies widely between fields (broadly corresponding to the distinction between "hard" and "soft" sciences). Mathematics, for instance, tends to focus almost entirely (in the short term, at least) on descriptive activity (e.g. determining the truth or falsity of various conjectures, solving problems, or proving theorems), although visionary (and prescriptivist) guidance (e.g. introducing a point of view, making an influential set of conjectures, identifying promising avenues of research, initiating a mathematical program, or finding the "right" definition for a mathematical concept, or the "right" set of axioms for a formal system) does play a vital role in the long-term development of the field.

The physical sciences are often presented to the public from a prescriptive standpoint, in that they are supposed to answer the question of why nature is the way we see it to be, and what causes a certain physical phenomenon to happen. However, in truth, many of the successful and tangible achievements of physics have come instead from the descriptive side of the field - finding out what the laws of nature are, and how specific physical systems will behave. The relationship between the prescriptive and descriptive sides of physics is roughly analogous to the relationship between causation and correlation in statistics; the latter can (and should) form a supporting

[1] In some fields, "descriptive" and "prescriptive" are referred to as "positive" and "normative" respectively.

foundation of evidence for the former, but an understanding of the latter does not necessarily entail a corresponding understanding of the former.

The prescriptive side of physics is extremely difficult to formalise properly, as one can see by the immense literature on philosophy of science; it is not easy at all to quantify the extent to which the answer to a "why?" or "what causes?" question is correct and intellectually satisfying.

In contrast, the descriptive side of physics, while perhaps less satisfying, is at least somewhat easier to formalise (though it is not without its own set of difficulties, such as the problem of defining precisely what a measurement or observation is, and how to deal with errors in the measurements or in the model). One way to do so is to take a computational complexity viewpoint, and view descriptive physics as an effort to obtain increasingly good upper bounds on the descriptive complexity (or *Kolmogorov complexity*) of the universe, or more precisely on the set of observations that we can make in the universe.

To give an example of this, consider a very simple set of observations, namely the orbital periods $T_1, \ldots, T_6$ of the six classical planets (Mercury, Venus, Earth, Mars, Jupiter, Saturn), and their distances $R_1, \ldots, R_6$ to the Sun (ignoring for now the detail that the orbits are not quite circular, but are instead essentially elliptical). To describe this data set, one could perform[2] the relevant set of observations, and obtain a list of twelve numbers $T_1, \ldots, T_6, R_1, \ldots, R_6$, which form a complete description of this data set.

On the other hand, if one is aware of *Kepler's third law*, one knows about the proportionality relationship

$$T_i^2 = cR_i^3$$

for some constant c and all $i = 1, \ldots, 6$. In that case, one can describe the entire data set by just seven numbers - the distances $R_1, \ldots, R_6$ and the constant c - together with Kepler's third law. This is a shorter description of the set, and so we have thus reduced the upper bound on the Kolmogorov complexity of the set. In this example, we have only shortened the length of the description by five numbers (minus the length required to state Kepler's law), but if one then adds in more planets and planet-like objects (e.g. asteroids, and also comets if one generalises Kepler's law to elliptical orbits), one sees the improvement in descriptive complexity become increasingly marked. In particular, the "one-time cost" of stating Kepler's law (and of stating the proportionally constant c) eventually becomes a negligible component of the

[2]For this exercise, we will ignore the issue of possible inaccuracies in measurement, or in the implicit physical assumptions used to perform such a measurement.

total descriptive complexity, when the range of applicability of the law becomes large. This is in contrast to superficially similar proposed laws such as the *Titius-Bode law*, which was basically restricted to the six classical planets and thus provided only a negligible saving in descriptive complexity.

Note that Kepler's law introduces a new quantity, c, to the explanatory model of the universe. This quantity increases the descriptive complexity of the model by one number, but this increase is more than offset by the decrease (of six numbers, in the classical case) caused by the application of the law. Thus we see the somewhat unintuitive fact that one can simplify one's model of the universe by adding parameters to it. However, if one adds a gratuitiously large number of such parameters to the model, then one can end up with a net increase in descriptive complexity, which is undesirable; this can be viewed as a formal manifestation of Occam's razor. For instance, if one had to add an *ad hoc* "fudge factor" F_i to Kepler's law to make it work,

$$T_i^2 = cR_i^3 + F_i,$$

with F_i being different for each planet, then the descriptive complexity of this model has in fact increased to thirteen numbers (e.g. one can specify $c, R_1, \ldots R_6$, and $F_1, \ldots, F_6$), together with the fudged Kepler's law, leading to a model with worse complexity[3] than the initial model of simply stating all the twelve observables $T_1, \ldots, T_6, R_1, \ldots, R_6$.

Note also that the additional parameters (such as c) introduced by such a law were not initially present in the previous model of the data set, and can only be measured through the law itself. This can give the appearance of circularity - Kepler's law relates times and radii of planets using a constant c, but the constant c can only be determined by applying Kepler's law. If there was only one planet in the data set, this law would indeed be circular (providing no new information on the orbital time and radius of the planet); but the power of the law comes from its uniform applicability among all planets. For instance, one can use data from the six classical planets to compute c, which can then be used to make predictions on, say, the orbital period of a newly discovered planet at a known distance to the sun. This may seem confusingly circular[4] from the prescriptive viewpoint - does the

[3]However, if this very same fudge factor F_i also appeared in laws that involved other statistics of the planet, e.g. mass, radius, temperature, etc. - then it can become possible again that such a law could act to decrease descriptive complexity when working with an enlarged data set that involves these statistics. Also, if the fudge factor is always small, then there is still some decrease in descriptive complexity coming from a saving in the most significant figures of the primary measurements T_i, R_i. So an analysis of an oversimplified data set, such as this one, can be misleading.

[4]One could use mathematical manipulation to try to eliminate such unsightly constants, for instance replacing Kepler's law with the (mathematically equivalent) assertion that $T_i^2/R_i^3 =$

constant c "cause" the relationship between period and distance, or vice versa? - but is perfectly consistent and useful from the descriptive viewpoint.

Note also that with this descriptive approach to Kepler's law, absolutely nothing has been said about the *causal* origins of the law. Of course, we now know that Kepler's law can be mathematically deduced from Newton's law of gravitation (which has a far greater explanatory power, and thus achieves a far greater reduction in descriptive complexity, than Kepler's laws, due to its much wider range of applicability). From a prescriptive viewpoint, this can be viewed as a partial explanation of Kepler's law, reducing the question to that of understanding the causal origins of Newton's law. When viewed in isolation, this may not be regarded as much of a reduction, as one is simply replacing one unexplained law with another; but when one takes into account that Newton's laws of classical mechanics can be used to derive hundreds of previously known classical laws besides Kepler's law, we see that Newtonian mechanics did in fact achieve a substantial reduction in the number of unexplained laws in physics. Thus we see that descriptive science can be used to reduce the magnitude of problems one faces in prescriptive science, although it cannot by itself be used to solve these problems entirely.

In modern physics, of course, we model the universe to be extremely large, extremely old, and to have structure both at very fine scales and very large scales. At first glance, this seems to massively increase the descriptive complexity of this model, in defiance of Occam's razor. However, these scale parameters in our model were not chosen gratuitously, but were the natural and consistent consequence of extrapolating from the known observational data using the known laws of physics. All known rival models of the universe that are significantly smaller in scale in either time or space require either that a large fraction of observational data be arbitrarily invalidated, or that the known laws of physics acquire an ad hoc set of fudge factors that emerge in some range of physical scenarios but not in others (in particular, these factors need to somehow disappear in all scenarios that can be directly observed). Either of these two "fixes" ends up leading to a much larger descriptive complexity for the universe than the standard model.

In some cases, the additional parameters introduced by a model to reduce the descriptive complexity are in fact unphysical - they cannot be computed, even in principle, from observation and from the laws of the model. A simple example is that of the potential energy of an object in classical physics. Experiments (e.g. measuring the amount of work needed to alter the state of an object) can measure the difference between the potential energy of

T_j^2/R_j^3 for all i, j, but this tends to lead to mathematically uglier laws and also does not lead to any substantial saving in descriptive complexity.

an object in two different states, but cannot compute[5] the potential energy itself. Indeed, one could add a fixed constant to the potential energy of all the possible states of an object, and this would not alter any of the physical consequences of the model. Nevertheless, the presence of such unphysical quantities can serve to reduce the descriptive complexity of a model (or at least to reduce the mathematical complexity, by making it easier to compute with the model), and can thus be desirable from a descriptive viewpoint, even though they are unappealing from a prescriptive one.

It is also possible to use mathematical abstraction to reduce the number of unphysical quantities in a model; for instance, potential energy could be viewed not as a scalar, but instead as a more abstract *torsor*. Again, these mathematical manipulations do not fundamentally affect the physical consequences of the model.

10.3. Honesty and Bayesian probability

Suppose you are shopping for some item X. You find a vendor V who is willing to sell X to you at a good price. However, you do not know whether V is honest (and thus selling you a genuine X), or dishonest (selling you a counterfeit X). How can one estimate the likelihood that V is actually honest?

One can try to model this problem using Bayesian probability. One can assign a prior probability p that V is honest (based, perhaps, on how trustworthy V looks, or on past experience with such vendors). However, one can update this prior probability p based on contextual information, such as the nature of the deal V is offering you, the way in which you got in contact with V, the venue in which V is operating in, and the past history of V (or the brand that V represents).

For instance, suppose V is offering you X at a remarkably low price Y - one which is almost "too good to be true". Specifically, this price might be so low that an honest vendor would find it very difficult to sell X profitably at this price, whereas a dishonest vendor could more easily sell a counterfeit X at the same price. Intuitively, this context should create a downward revision on one's probability estimate that V is honest. Indeed, if we let a be the conditional probability

$$a := \mathbf{P}(V \text{ sells at } Y | V \text{ is honest})$$

and b be the probability

[5]Amusingly, in special relativity, the potential energy does actually become physically measurable, thanks to Einstein's famous equation $E = mc^2$, but this does not detract from the previous point. Other examples of non-physical quantities that are nevertheless descriptively useful include the wave function in quantum mechanics, or gauge fields in gauge theory.

$$b := P(V \text{ sells at } Y | V \text{ is dishonest})$$

then after a bit of computation using Bayes' theorem, we find that

$$P(V \text{ is honest} | V \text{ sells at } Y) = \frac{ap}{ap + b(1-p)}. \tag{10.2}$$

the right-hand side can be rearranged as

$$p - \frac{(b-a)p(1-p)}{ap + b(1-p)}.$$

Thus we do indeed see that if $b > a$, then the probability that V is honest is revised downwards from p (and conversely if $b < a$, then we revise the probability that V is honest upwards).

In a similar fashion, if V has invested in a substantial storefront presence, which would make it difficult (or at least expensive) for V to quickly disappear in case of customer complaints about X, then the same analysis increases the probability that V is honest, since it is unlikely that a dishonest vendor would make such an investment, instead preferring a more mobile "fly by night" operation. Or in the language of the above Bayesian analysis: the analogue of a is large, and the analogue of b is small.

One can also take V's past sale history into account. Suppose that one knows that V has already sold N copies of X without any known complaint. If we make the somewhat idealistic assumptions that an honest vendor would not cause any complaints, and each sale by a dishonest vendor has a probability ε of causing a complaint (with the probability of complaint being independent from sale to sale), then in the notation of the previous analysis, we have $a = 1$ and $b = (1-\varepsilon)^N$. As N gets large, b tends exponentially to zero, and this causes the posterior probability that V is honest to tend exponentially to 1, as can be seen by the formula (10.2). This analysis can help explain the power of large corporate brands, which have a very long history of sales, and thus (assuming, of course, that their prior reputation is strong) have a significant advantage over smaller competitors in that consumers generally entrust them to guarantee a certain minimum level of quality. (Conversely, smaller businesses can take more risks, and can thus sometimes offer levels of quality significantly higher than that of a safe corporate brand.)

A similar analysis can be applied to non-commercial settings, such as the leak of some purportedly genuine document. If one has an anonymous leak of only a single document, then it can be quite difficult to determine whether the document is genuine or not, as it is entirely possible to forge a single document that passes for genuine under superficial scrutiny. However,

if there is a leak of N documents for a large value of N, and no glaring inaccuracies or contradictions have been found in any of these documents, then the probability that the documents are largely genuine converges quite rapidly to one, because the difficulty of forging N documents without any obvious slip-ups increases exponentially with N.

It is important to note, however, that Bayesian analysis is only as strong as the assumptions that underlie it. In the above analysis that a long history of sales without complaint increases the probability that the vendor is honest, an important assumption was made that each sale by a dishonest vendor had an independent probability of triggering a complaint. However, this assumption can fail in some key situations, most notably when X is a financial product, and the vendor V could potentially be running a pyramid scheme. In such schemes, there are essentially no complaints from customers for most of the lifetime of the scheme, but then there is a catastrophic collapse at the very end of the scheme. As such, a past history of satisfied customers does not in fact increase the probability that V is honest in this case. (Another thing to note is that pyramid schemes, by their nature, grow exponentially in time, and so one is statistically much more likely to come in contact with a pyramid scheme when it is large and near the end of its lifespan, than when it is small and still some way from collapsing.)

Bibliography

[Al2011] J. M. Aldaz, *The weak type* $(1,1)$ *bounds for the maximal function associated to cubes grow to infinity with the dimension*, Ann. of Math. (2) **173** (2011), no. 2, 1013-1023.

[Al1974] F. Alexander, *Compact and finite rank operators on subspaces of* l^p, Bull. London Math. Soc. **6** (1974), 341-342.

[AmGe1973] W. O. Amrein, V. Georgescu, *On the characterization of bound states and scattering states in quantum mechanics*, Helv. Phys. Acta **46** (1973/74), 635-658.

[Ba1966] A. Baker, *Linear forms in the logarithms of algebraic numbers. I*, Mathematika. A Journal of Pure and Applied Mathematics **13** (1966), 204-216.

[Ba1967] A. Baker, *Linear forms in the logarithms of algebraic numbers. II*, Mathematika. A Journal of Pure and Applied Mathematics **14** (1966), 102-107.

[Ba1967b] A. Baker, *Linear forms in the logarithms of algebraic numbers. III*, Mathematika. A Journal of Pure and Applied Mathematics **14** (1966), 220-228.

[BoSo1978] C. Böhm, G. Sontacchi, *On the existence of cycles of given length in integer sequences like* $x_{n+1} = x_{n/2}$ *if* x_n *even, and* $x_{n+1} = 3x_n + 1$ *otherwise*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. **64** (1978), no. 3, 260-264.

[BoChLoSoVe2008] C. Borgs, J. Chayes, L. Lovász, V. Sós, K. Vesztergombi, *Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing*, Adv. Math. **219** (2008), no. 6, 1801-1851.

[Bo1985] J. Bourgain, *Estimations de certaines fonctions maximales*, C. R. Acad. Sci. Paris Sér. I Math. **301** (1985), no. 10, 499-502.

[Bo1991] J. Bourgain, *Besicovitch type maximal operators and applications to Fourier analysis*, Geom. Funct. Anal. **1** (1991), no. 2, 147-187.

[Bo2005] J. Bourgain, *Estimates on exponential sums related to the Diffie-Hellman distributions*, Geom. Funct. Anal. **15** (2005), no. 1, 1-34.

[BoSaZi2011] J. Bourgain, P. Sarnak, T. Ziegler, *Disjointness of Mobius from horocycle flows*, preprint.

[BrGrGuTa2010] E. Breuillard, B. Green, R. Guralnick, T. Tao, *Strongly dense free subgroups of semisimple algebraic groups*, preprint.

[CaRuVe1988] A. Carbery, J. Rubio de Francia, L. Vega, *Almost everywhere summability of Fourier integrals*, J. London Math. Soc. (2) **38** (1988), no. 3, 513-524.

[Ca1966] L. Carleson, *On convergence and growth of partial sums of Fourier series*, Acta Mathematica **116** (1966), 135-157.

[Ca1813] A. L. Cauchy, *Recherches sur les nombres*, J. École Polytech. **9** (1813), 99-116.

[Ce1964] A. V. Cernavskii, *Finite-to-one open mappings of manifolds*, Mat. Sb. (N.S.) **65** (1964), 357-369.

[Ch2003] M. Chang, *Factorization in generalized arithmetic progressions and applications to the Erdos-Szemerédi sum-product problems*, Geom. Funct. Anal. **13** (2003), no. 4, 720-736.

[Ch1885] M. Chasles, *Traité des sections coniques*, Gauthier-Villars, Paris, 1885.

[Ch2008] M. Christ, *Quasi-extremals for a Radon-like transform*, preprint. `www.math.berkeley.edu/ mchrist/Papers/quasiextremal.pdf`

[Co2007] A. Comech, *Cotlar-Stein almost orthogonality lemma*, preprint. `www.math.tamu.edu/ comech/papers/CotlarStein/CotlarStein.pdf`

[CoCoGr2008] B. Conrad, K. Conrad, R. Gross, *Prime specialization in genus* 0, Trans. Amer. Math. Soc. **360** (2008), no. 6, 2867-2908.

[Co1955] M. Cotlar, *A combinatorial inequality and its application to L^2 spaces*, Math. Cuyana **1** (1955), 41-55.

[Da1935] H. Davenport, *On the addition of residue classes*, J. London Math. Soc. **10** (1935), 30–32.

[de1981] M. de Guzmán, Real variable methods in Fourier analysis. North-Holland Mathematics Studies, 46. Notas de Matemtica [Mathematical Notes], 75. North-Holland Publishing Co., Amsterdam-New York, 1981.

[de2006] T. de la Rue, 2*-fold and* 3*-fold mixing: why* 3*-dot-type counterexamples are impossible in one dimension*, Bull. Braz. Math. Soc. (N.S.) **37** (2006), no. 4, 503-521.

[De1971] F. Delmer, *Sur la somme de diviseurs* $\sum_{k\leq x} d[f(k)]^s$, C. R. Acad. Sci. Paris Sér. A-B **272** (1971), A849-A852.

[DeFoMaWr2010] S. Dendrinos, M. Folch-Gabayet, J. Wright, *An affine-invariant inequality for rational functions and applications in harmonic analysis*, Proc. Edinb. Math. Soc. (2) **53** (2010), no. 3, 639-655.

[Ei1969] D. Éidus, *The principle of limiting amplitude*, Uspehi Mat. Nauk **24** (1969), no. 3(147), 91-156.

[ElSz2012] G. Elek, B. Szegedy, *A measure-theoretic approach to the theory of dense hypergraphs*, Adv. Math. **231** (2012), no. 3-4, 1731-1772.

[ElSh2011] G. Elekes, M. Sharir, *Incidences in three dimensions and distinct distances in the plane*, Combin. Probab. Comput. **20** (2011), no. 4, 571-608.

[ElObTa2010] J. Ellenberg, R. Oberlin, T. Tao, *The Kakeya set and maximal conjectures for algebraic varieties over finite fields*, Mathematika **56** (2010), no. 1, 1-25.

[ElTa2011] C. Elsholtz, T. Tao, *Counting the number of solutions to the Erdos-Straus equation on unit fractions*, preprint.

[En1973] P. Enflo, *A counterexample to the approximation problem in Banach spaces*, Acta Math. **130** (1973), 309-317.

[En1978] V. Enss, *Asymptotic completeness for quantum mechanical potential scattering. I. Short range potentials*, Comm. Math. Phys. **61** (1978), no. 3, 285-291.

[Er1952] P. Erdős, *On the sum* $\sum_{k=1}^{x} d(f(k))$, J. London Math. Soc. **27** (1952), 7-15.

[Er1979] P. Erdős, *Some unconventional problems in number theory*, Journées Arithmétiques de Luminy (Colloq. Internat. CNRS, Centre Univ. Luminy, Luminy, 1978), pp. 73-82, Astérisque, 61, Soc. Math. France, Paris, 1979.

[Ka1940] P. Erdos, M. Kac, *The Gaussian Law of Errors in the Theory of Additive Number Theoretic Functions*, American Journal of Mathematics **62** (1940), 738-742.

[Fe1971] C. Fefferman, *The multiplier problem for the ball*, Ann. of Math. (2) **94** (1971), 330-336.

[Fe1995] C. Fefferman, *Selected theorems by Eli Stein*, Essays on Fourier analysis in honor of Elias M. Stein (Princeton, NJ, 1991), 135, Princeton Math. Ser., 42, Princeton Univ. Press, Princeton, NJ, 1995.

[FeSt1972] C. Fefferman, E. Stein, *H^p spaces of several variables*, Acta Math. **129** (1972), no. 3-4, 137-193.

[Fu1977] H. Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. **31** (1977), 204–256.

[Ga1981] L. Garner, *On the Collatz $3n+1$ algorithm*, Proc. Amer. Math. Soc. **82** (1981), no. 1, 19-22.

[Ge1934] A. Gelfond, *Sur le septiéme Probléme de D. Hilbert*,Comptes Rendus Acad. Sci. URSS Moscou **2** (1934), 1–6.

[Go2008] W. T. Gowers, *Quasirandom groups*, Combin. Probab. Comput. **17** (2008), no. 3, 363-387.

[Gr2008] A. Granville, *Smooth numbers: computational number theory and beyond*, Algorithmic number theory: lattices, number fields, curves and cryptography, 267323, Math. Sci. Res. Inst. Publ., 44, Cambridge Univ. Press, Cambridge, 2008.

[Gr1970] G. Greaves, *On the divisor-sum problem for binary cubic forms*, Acta Arith. **17** (1970) 1-28.

[GrRu2005] B. Green, I. Ruzsa, *Sum-free sets in abelian groups*, Israel J. Math. **147** (2005), 157-188.

[GrTa2012] B. Green, T. Tao, *The Möbius function is strongly orthogonal to nilsequences*, Ann. of Math. (2) **175** (2012), no. 2, 541-566.

[Gr1955] A. Grothendieck, *Produits tensoriels topologiques et espaces nucléaires*, Mem. Amer. Math. Soc. **1955** (1955), no. 16, 140 pp.

[Gu2011] C. Gunn, *On the Homogeneous Model Of Euclidean Geometry*, AGACSE (2011)

[GuKa2010] L. Guth, N. Katz, *On the Erdos distinct distance problem in the plane*, preprint.

[Gu1988] R. Guy, *The Strong Law of Small Numbers*, American Mathematical Monthly **95** (1988), 697-712.

[Ha2010] Y. Hamidoune, *Two Inverse results*, preprint. `arXiv:1006.5074`

[HaRa1917] G. H. Hardy, S. Ramanujan, *The normal number of prime factors of a number*, Quarterly Journal of Mathematics **48** (1917), 76-92.

[He1983] J. Heintz, *Definability and fast quantifier elimination over algebraically closed fields*, Theoret. Comput. Sci. **24** (1983), 239–277.

[Ho1963] C. Hooley, *On the number of divisors of a quadratic polynomial*, Acta Math. **110** (1963), 97-114.

[Ho1991] B. Host, *Mixing of all orders and pairwise independent joinings of systems with singular spectrum*, Israel J. Math. **76** (1991), no. 3, 289-298.

[Hr2012] E. Hrushovski, *Stable group theory and approximate subgroups*, J. Amer. Math. Soc. **25** (2012), no. 1, 189-243.

[Hu2004] D. Husemöller, Elliptic curves. Second edition. With appendices by Otto Forster, Ruth Lawrence and Stefan Theisen. Graduate Texts in Mathematics, 111. Springer-Verlag, New York, 2004.

[IoRoRu2011] A. Iosevich, O. Roche-Newton, M. Rudnev, *On an application of Guth-Katz theorem*, preprint.

[Ka1986] I. Kátai, A remark on a theorem of H. Daboussi. Acta Math. Hungar. **47** (1986), no. 1-2, 223-225.

[Ka1984] S. Kalikow, *Twofold mixing implies threefold mixing for rank one transformations*, Ergodic Theory Dynam. Systems **4** (1984), no. 2, 237-259.

[Ka1965] T. Kato, *Wave operators and similarity for some non-selfadjoint operators*, Math. Ann. **162** (1965/1966), 258-279.

[Ke1964] J. H. B. Kemperman, *On products of sets in locally compact groups*, Fund. Math. **56** (1964), 51-68.

[KnSt1971] A. Knapp, E. Stein, *Intertwining operators for semisimple groups*, Ann. of Math. (2) **93** (1971), 489-578.

[Kn1953] M. Kneser, *Abschätzungen der asymptotischen Dichte von Summenmengen*, Math. Z **58** (1953), 459–484.

[KrLa2003] I. Krasikov, J. Lagarias, *Bounds for the* $3x+1$ *problem using difference inequalities*, Acta Arith. **109** (2003), no. 3, 237-258.

[KrRa2010] S. Kritchman, R. Raz, *The surprise examination paradox and the second incompleteness theorem*, Notices Amer. Math. Soc. **57** (2010), no. 11, 1454-1458.

[La2009] J. Lagarias, *Ternary expansions of powers of* 2., J. Lond. Math. Soc. **79** (2009), no. 3, 562-588.

[La1989] B. Landreau, *A new proof of a theorem of van der Corput*, Bull. London Math. Soc. **21** (1989), no. 4, 366-368.

[Le1978] F. Ledrappier, *Un champ markovien peut être d'entropie nulle et mélangeant*, C. R. Acad. Sci. Paris Sér. A-B **287** (1978), no. 7, A561-A563.

[LeMa2005] G. Leonardi, S. Masnou, *On the isoperimetric problem in the Heisenberg group* $\mathbb{H}^n$, Ann. Mat. Pura Appl. (4) **184** (2005), no. 4, 533553.

[Li1973] W. Littman, $L^p - L^q$*-estimates for singular integral operators arising from hyperbolic equations*, Partial differential equations (Proc. Sympos. Pure Math., Vol. XXIII, Univ. California, Berkeley, Calif., 1971), pp. 479481. Amer. Math. Soc., Providence, R.I., 1973.

[Lo1975] P. Loeb, *Conversion from nonstandard to standard measure spaces and applications in probability theory*, Trans. Amer. Math. Soc. **211** (1975), 113-122.

[LoSz2006] L. Lovász, B. Szegedy, *Limits of dense graph sequences*, J. Combin. Theory Ser. B **96** (2006), no. 6, 933-957.

[Ma1953] A. M. Macbeath, *On measure of sum sets. II. The sum-theorem for the torus*, Proc. Cambridge Philos. Soc. **49**, (1953), 40-43.

[MaHu2008] C. R. MacCluer, A. Hull, *A short proof of the Fredholm alternative*, Int. J. Pure Appl. Math. **45** (2008), no. 3, 379-381.

[Ma2010] K. Maples, *Singularity of Random Matrices over Finite Fields*, preprint. `arXiv:1012.2372`

[Ma2012] L. Matthiesen, *Correlations of the divisor function*, Proc. Lond. Math. Soc. **104** (2012), 827-858.

[Ma1974] B. Maurey, *Théorèmes de factorisation pour les opérateurs linéaires à valeurs dans les espaces L^p*, With an English summary. Astrisque, No. 11. Socit Mathmatique de France, Paris, 1974 ii+163 pp.

[Mc1995] J. McKee, *On the average number of divisors of quadratic polynomials*, Math. Proc. Cambridge Philos. Soc. **117** (1995), no. 3, 389-392.

[Mc1997] J. McKee, 'A note on the number of divisors of quadratic polynomials. Sieve methods, exponential sums, and their applications in number theory' (Cardiff, 1995), 275–281, *London Math. Soc. Lecture Note Ser.*, 237, Cambridge Univ. Press, Cambridge, 1997.

[Mc1999] J. McKee, 'The average number of divisors of an irreducible quadratic polynomial', *Math. Proc. Cambridge Philos. Soc.* **126** (1999), no. 1, 17–22.

[Mi1964] J. Milnor, *On the Betti numbers of real varieties*, Proc. Amer. Math. Soc. **15** (1964), 275-280.

[MoSeSo1992] G. Mockenhaupt, A. Seeger, C. Sogge, *Wave front sets, local smoothing and Bourgain's circular maximal theorem*, Ann. of Math. (2) **136** (1992), no. 1, 207218.

[Mo2003] R. Monti, *Brunn-Minkowski and isoperimetric inequality in the Heisenberg group*, Ann. Acad. Sci. Fenn. Math. **28** (2003), no. 1, 99-109.

[Ni1970] E. Nikishin, *Resonance theorems and superlinear operators*, Uspehi Mat. Nauk **25** (1970), no. 6 (156), 129191.

[Ob1992] D. Oberlin, *Multilinear proofs for two theorems on circular averages*, Colloq. Math. **63** (1992), no. 2, 187-190.

[OlPe1949] I. G. Petrovskii, O. A. Oleinik, *On the topology of real algebraic surfaces*, Izvestiya Akad. Nauk SSSR. Ser. Mat. **13** (1949), 389-402.

[Pi1986] G. Pisier, *Factorization of operators through $L_{p,\infty}$ or $L_{p,1}$ and noncommutative generalizations*, Math. Ann. **276** (1986), no. 1, 105-136.

[Po1974] J. M. Pollard, *A generalisation of the theorem of Cauchy and Davenport*, J. London Math. Soc. **8** (1974), 460-462.

[Pr2007] C. Procesi, Lie groups. An approach through invariants and representations. Universitext. Springer, New York, 2007.

[Ra1939] D. Raikov, *On the addition of point-sets in the sense of Schnirelmann*, Rec. Math. [Mat. Sbornik] N.S. **5**, (1939), 425-440.

[RoTa2011] I. Rodnianski, T. Tao, *Effective limiting absorption principles, and applications*, preprint.

[Ro1949] V. A. Rohlin, *On endomorphisms of compact commutative groups*, Izvestiya Akad. Nauk SSSR. Ser. Mat. **13**, (1949), 329-340.

[Ro1953] K. F. Roth, *On certain sets of integers*, J. London Math. Soc. **28** (1953), 245–252.

[Ro1955] K. F. Roth, *Rational approximations to algebraic numbers*, Mathematika **2** (1955), 1-20.

[Ru1969] D. Ruelle, *A remark on bound states in potential-scattering theory*, Nuovo Cimento A **61** (1969), 655-662.

[Ru1992] I. Ruzsa, *A concavity property for the measure of product sets in groups*, Fund. Math. **140** (1992), no. 3, 247-254.

[RuSz1978] I. Ruzsa, E. Szemerédi, *Triple systems with no six points carrying three triangles*, Colloq. Math. Soc. J. Bolyai **18** (1978), 939–945.

[Ra2002] J. Saint Raymond, *Local inversion for differentiable functions and the Darboux property*, Mathematika **49** (2002), 141-158.

[Sc1995] J. Schmid, *On the affine Bézout inequality*, Manuscripta Mathematica **88** (1995), Number 1, 225–232.

[Sc1989] K. Schmidt-Göttsch, *Polynomial bounds in polynomial rings over fields*, J. Algebra **125** (1989), no. 1, 164-180.

[Sc1934] T. Schneider, *Transzendenzuntersuchungen periodischer Funktionen. I*, J. reine angew. Math. **172** (1934), 65–69.

[Se1969] J. P. Serre, *Travaux de Baker*, Séminaire Bourbaki, exp. 368 (1969–1970), 73–86.

[Si1921] C. L. Siegel, *Approximation algebraischer Zahlen*, Mathematische Zeitschrift **10** (1921), 173-213.

[Side2005] J. Simons, B. de Weger, *Theoretical and computational bounds for m-cycles of the $3n+1$-problem*, Acta Arith. **117** (2005), no. 1, 51-70.

[SoTa2011] J. Solymosi, T. Tao, *An incidence theorem in higher dimensions*, preprint.

[So2010] K. Soundararajan, *Math249A Fall 2010: Transcendental Number Theory*, lecture notes available at `math.stanford.edu/ ksound/TransNotes.pdf`. Transcribed by Ian Petrow.

[St1956] E. M. Stein, *Interpolation of linear operators*, Trans. Amer. Math. Soc. **83** (1956), 482-492.

[St1961] E. M. Stein, *On limits of seqences of operators*, Ann. of Math. (2) **74** (1961) 140-170.

[St1976] E. M. Stein, *Maximal functions. I. Spherical means*, Proc. Nat. Acad. Sci. U.S.A. **73** (1976), no. 7, 2174-2175.

[St1982] E. M. Stein, *The development of square functions in the work of A. Zygmund*, Bull. Amer. Math. Soc. (N.S.) **7** (1982), no. 2, 359-376.

[St1993] E. M. Stein, Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals. With the assistance of Timothy S. Murphy. Princeton Mathematical Series, 43. Monographs in Harmonic Analysis, III. Princeton University Press, Princeton, NJ, 1993.

[StSt1983] E. M. Stein, J.-O. Strömberg, *Behavior of maximal functions in R^n for large n*, Ark. Mat. **21** (1983), no. 2, 259-269.

[St1978] P. Steiner, *A theorem on the Syracuse problem*, Proceedings of the Seventh Manitoba Conference on Numerical Mathematics and Computing (Univ. Manitoba, Winnipeg, Man., 1977), pp. 553559, Congress. Numer., XX, Utilitas Math., Winnipeg, Man., 1978.

[St2010] B. Stovall, *Endpoint $L^p - L^q$ bounds for integration along certain polynomial curves*, J. Funct. Anal. **259** (2010), no. 12, 3205-3229.

[St1891] E. Study, *Von den bewegungen und umlegungen*, Mathematische Annalen, **39** (1891), 441-566.

[Sw1962] R. Swan, *Factorization of polynomials over finite fields*, Pacific J. Math., **12**, 1962, 1099-1106.

[Sz1975] E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27** (1975), 299–345.

[Sz1978] E. Szemerédi, *Regular partitions of graphs*, Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976), Colloq. Internat. CNRS, 260, Paris: CNRS, pp. 399-401.

[Ta2007] T. Tao, *A correspondence principle between (hyper)graph theory and probability theory, and the (hyper)graph removal lemma*, J. Anal. Math. **103** (2007), 1-45.

[Ta2008] T. Tao, Structure and randomness: pages from year one of a mathematical blog, American Mathematical Society, Providence RI, 2008.

[Ta2009] T. Tao, Poincaré's Legacies: pages from year two of a mathematical blog, Vol. I, American Mathematical Society, Providence RI, 2009.

[Ta2009b] T. Tao, Poincaré's Legacies: pages from year two of a mathematical blog, Vol. II, American Mathematical Society, Providence RI, 2009.

[Ta2010] T. Tao, An epsilon of room, Vol. I, American Mathematical Society, Providence RI, 2010.

[Ta2010b] T. Tao, An epsilon of room, Vol. II, American Mathematical Society, Providence RI, 2010.

[Ta2011] T. Tao, An introduction to measure theory, American Mathematical Society, Providence RI, 2011.

[Ta2011b] T. Tao, Higher order Fourier analysis, American Mathematical Society, Providence RI, 2011.

[Ta2011c] T. Tao, Topics in random matrix theory, American Mathematical Society, Providence RI, 2011.

[Ta2011d] T. Tao, Compactness and contradiction, American Mathematical Society, Providence RI, 2011.

[Ta2012] T. Tao, Hilbert's fifth problem and related topics, in preparation.

[Ta2012b] T. Tao, Noncommutative sets of small doubling, preprint.

[TaVu2006] T. Tao, V. Vu, Additive combinatorics, Cambridge University Press, 2006.

[Te1976] R. Terras, *A stopping time problem on the positive integers*, Acta Arith. 30 (1976), no. 3, 241-252.

[Th1965] R. Thom, *Sur l'homologie des variétés algébriques réelles*, 1965 Differential and Combinatorial Topology (A Symposium in Honor of Marston Morse) pp. 255-265 Princeton Univ. Press, Princeton, N.J.

[Th1909] A. Thue, *Über Annäherungswerte algebraischer Zahlen*, Journal fr die reine und angewandte Mathematik **135** (1909), 284-305.

[Uu2010] O. Uuye, *A simple proof of the Fredholm alternative*, preprint. `arXiv:1011.2933`

[Va1966] J. Väisälä, *Discrete open mappings on manifolds*, Ann. Acad. Sci. Fenn. Ser. A I **392** (1966), 10 pp.

[va1939] J. G. van der Corput, *Une inégalité relative au nombre des diviseurs*, Nederl. Akad. Wetensch., Proc. **42** (1939), 547-553.

[vdW1927] B.L. van der Waerden, *Beweis einer Baudetschen Vermutung*, Nieuw. Arch. Wisk. **15** (1927), 212-216.

[Wa1836] M. L. Wantzel, *Recherches sur les moyens de reconnaître si un Probléme de Géométrie peut se résoudre avec la règle et le compas*, J. Math. pures appliq. **1** (1836), 366–37.

Index